REPORT RESUMES

ED 011 629 AC 000 534

CRITERIA IN LEARNING RESEARCH, REPORT ON A CONFERENCE (WASHINGTON UNIVERSITY, 1966).

BY- WIENTGE, KING M. DUBOIS, PHILIP H.
WASHINGTON UNIV., ST. LOUIS, UNIVERSITY COLL.
REPORT NUMBER WUCRP-D PUB DATE 66
EDRS PRICE MF-$0.18 HC-$2.96 74P.

DESCRIPTORS- *ADULT LEARNING, *LEARNING, *CRITERIA, *MEASUREMENT, *RESEARCH METHODOLOGY, RESEARCH, PERFORMANCE, TABLES (DATA), EVALUATION TECHNIQUES, STATISTICAL DATA, PROGRAMED INSTRUCTION, PREDICTIVE ABILITY (TESTING), EDUCATIONAL ENVIRONMENT, ACHIEVEMENT, ST. LOUIS

THE TOPICS OF EIGHT CONFERENCE PAPERS INCLUDE (1) THE PROBLEM OF DIFFERENTIATING EFFECTS OF SPECIFIC INSTRUCTION FROM EFFECTS OF OTHER INFLUENCES (GROWTH, ENVIRONMENT, AND SELF-INSTRUCTION), (2) CRITERIA FOR MEASURING CHANGE IN PROFICIENCY, AND (3) WAYS OF RELATING SUCH CHANGE TO OUTSIDE VARIABLES SUCH AS MEASURABLE CHARACTERISTICS OF LEARNERS AND INSTRUCTORS, AND METHODOLOGY. THE ISSUE OF ACCEPTABLE CRITERIA (PERFORMANCE MEASURES) OF DEGREES OF LEARNING DURING PRACTICE IS DISCUSSED. SUGGESTIONS ARE OUTLINED FOR IMPROVING EXPERIMENTATION BY ASSEMBLING ALL APPROPRIATE MEASUREMENTS AND DATA, USING ORDERED HYPOTHESES, AND TREATING EXPERIMENTAL CLASSROOMS AS SINGLE SUBJECTS. THE BROMWOODS RESIDENTIAL CENTER STUDY OF WASHINGTON UNIVERSITY DESCRIBES THE FAILURE OF RESIDENTIAL STUDENTS TO SIGNIFICANTLY SURPASS ADULT EVENING CLASSES IN OBJECTIVE KNOWLEDGE (BEGINNING PSYCHOLOGY) OR IN IMPROVED MEASURED ATTITUDE. ALSO, DIFFICULTIES IN APPLICATION OF CRITERIA IN NAVAL MAINTENANCE TRAINING RESEARCH ARE DOCUMENTED, AND CORRECTIVE PROCEDURES ARE INDICATED. A PAPER ON THE CURVILINEAR RELATIONSHIP BETWEEN KNOWLEDGE AND TEST PERFORMANCE ARGUES FOR FINAL EXAMINATIONS AS THE BEST EXISTING INDICANT OF LEARNING. OTHER PAPERS OUTLINE A FRAMEWORK FOR STUDYING CRITERION MEASURES AND THEIR GENERALIZABILITY ACROSS SAMPLES, SITUATIONS, AND CONTEXTS, AND EVALUATE PROGRAMED INSTRUCTION AMONG NAVAL TRAINEES AS A PREDICTOR OF CLASSROOM LEARNING. THE DOCUMENT INCLUDES TABLES, FIGURES, AND REFERENCES. (LY)

ED011629

WASHINGTON UNIVERSITY
UNIVERSITY COLLEGE AND THE DEPARTMENT OF PSYCHOLOGY
Saint Louis, Missouri 63130

CRITERIA IN LEARNING RESEARCH

A Report on a Conference Held at the
Bromwoods Residential Center
of Washington University
Sponsored by the Department of Psychology
under Office of Naval Research Contract Nonr 816(14) and the
Division of Research and Development of University College

EDITED BY:

King M. Wientge

and

Philip H. DuBois

UNIVERSITY COLLEGE RESEARCH PUBLICATIONS

Number 9

1966

AC 000 534

UNIVERSITY COLLEGE

Officers of Government

Thomas H. Eliot, A.B., LL.B., LL.D.	Chancellor of the University
George Edward Pake, Ph.D.	Provost

Lynn W. Eley, Ph.D.	Dean of University College and Summer School
James K. Lahr, M.A.Ed.	Associate Dean
John B. Ervin, Ed.D.	Associate Dean and Director of the Summer School
Frederick J. Thumin, Ph.D.	Director, Adult Counseling Service
Myron A. Spohrer, A.M.	Registrar-Counselor
Kingsley M. Wientge, Ed.D.	Director Division of Research and Development
Malcolm Van Deursen, M.A.Ed.	Director, Conferences and Short Courses
Harry J. Gaffney, M.S.	Assistant Registrar-Counselor
Charles Thomas, A.M.	Director of Special Projects
Leonard Zwieg, M.A.	Director, Community Leadership Project
Mrs. Jean Pennington, M.A.Ed.	Coordinator Continuing Education for Women
Donald R. Schuster, M.S.	Counselor

J. George Robinson, Ph.D.	Associate Dean, Graduate School of Business Administration
Willard P. Armstrong, Ph.D.	Assistant Dean of the School of Engineering
Kelvin Ryals, M.A.Ed.	Administrative Assistant Graduate Institute of Education

Research Consultant

Philip H. DuBois	Professor of Psychology

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

WASHINGTON UNIVERSITY

Saint Louis, Missouri

CRITERIA IN LEARNING RESEARCH

A Report on a Conference Held at the
Bromwoods Residential Center
of Washington University
Sponsored by the Department of Psychology
under Office of Naval Research Contract Nonr 816(14) and the
Division of Research and Development of University College

Edited by:

King M. Wientge

and

Philip H. DuBois

UNIVERSITY COLLEGE RESEARCH PUBLICATIONS

Number 9

1966

FOREWORD

The Department of Psychology of Washington University under Contract Nonr 816(14) with the Office of Naval Research has been conducting research at the Naval Air Station, Memphis, Tennessee, on learning in classroom situations. On the campus of Washington University, University College and the Department of Psychology have been jointly conducting research in classroom settings on factors related to the academic achievement of adult students.

It seemed both timely and fruitful to conduct a joint conference, bringing together researchers from the two sources and a group of consultants from psychology and education.

The papers presented at the conference appeared to be of sufficient interest to merit distribution to individuals and organizations working in the field of education at the adult level.

Philip H. DuBois
King M. Wientge

CONTENTS

EARLY HISTORY OF THE APPRAISAL OF LEARNING
Philip H. DuBois 1

THE CRITERION PROBLEM IN LEARNING RESEARCH
Marion E. Bunch 4

CRITERIA IN CLASSROOM RESEARCH
Ellis B. Page. 11

RESEARCH WITH "NORMAL" ADULT POPULATIONS : THE BROMWOODS STUDY
King M. Wientge 22

ABILITY TO LEARN OPERATIONAL EQUIPMENT AND SYSTEMS AS A CRITERION IN LEARNING RESEARCH
G. Douglas Mayo 25

THE CURVILINEAR RELATIONSHIP BETWEEN KNOWLEDGE AND TEST PERFORMANCE: FINAL EXAMINATION AS THE BEST INDICANT OF LEARNING
Ronald P. Carver 32

THE CRITERION PROBLEM
Winton H. Manning. 50

PROGRAMMED INSTRUCTION AS A PREDICTOR OF CLASSROOM LEARNING
James L. Wardrop. 61

CONFERENCE PARTICIPANTS

Dr. J. R. Berkshire, U. S. Naval School of Aviation, Pensacola

Dr. Robert Buckhout, Washington University

Dr. Marion E. Bunch, Washington University

Mr. James R. Burmeister, Washington University

Dr. Ronald P. Carver, Washington University

Dr. Philip H. DuBois, Washington University

Dr. James M. Dunlap, University City Public Schools

Dr. Lynn W. Eley, Washington University

Mr. Harry J. Gaffney, Washington University

Mr. Edward V. Hackett, Memphis State University

Dr. Earl I. Jones, U. S. Naval Personnel Research Activity, San Diego

Dr. Winton H. Manning, Texas Christian University

Dr. G. Douglas Mayo, Naval Air Station, Memphis

Dr. Ellis B. Page, University of Connecticut

Mr. Edward N. Peters, Washington University

Dr. Carl J. Spies, Kent State University

Mr. Myron A. Spohrer, Washington University

Dr. Frederick J. Thumin, Washington University

Mr. James L. Wardrop, Washington University

Dr. King M. Wientge, Washington University

Dr. Richard H. Willis, Washington University

EARLY HISTORY OF THE APPRAISAL OF LEARNING

Philip H. DuBois

As the theme of this conference we have proposed the question of the distinction, if any, between proficiency on the one hand and the results of education or training on the other. First of all, it is easy to see that confusion is possible. Educators with students proficient in the area in which they are instructing have a tendency to assume credit for their students' performance. A pupil reflects glory on the teacher because it is the teacher in effect who has molded the student and has given him his status.

In the evaluation of attempts to differentiate between proficiency on the one hand and the results of training on the other, I see three historical stages.

In the Western world the first stage began among the ancients and continued to the year 1219. Schools were common in the ancient world, but examinations, as far as we know, were unknown. It apparently was easy for teachers of Greece and Rome to see that proficiency in basic skills increased concurrently with educational activities. Hence, proficiency as a product of instruction was assumed.

The earliest formal examinations in Europe were apparently those instituted at the University of Bologna early in the thirteenth century. Prior to the granting of a degree, the oral test of the competence of the student was conducted in private, followed by a public examination which was essentially a formality. For hundreds of years, university examinations continued as exclusively oral tests of proficiency.

One of the newer universities, Louvain, instituted competitive examinations, at least as early as 1441 which was sixteen years after its founding. Candidates were graded in four classes: "rigorosi" (honor men), "transibiles" (satisfactory), "gratiosi" (charity passes), and a fourth class of failures. It was actually the examination system at Louvain which helped to establish its outstanding academic reputation. Perhaps one of the reasons why written examinations were slow to develop was the fact that the writing materials available in the ancient world and in Medieval Europe were awkward or expensive or both. Taking a test on clay in cuneiform or on a wax tablet in Latin might have been possible theoretically but apparently no one had that idea for the measurement of proficiency. Both papyrus and parchment were expensive. Paper found its way to Europe through the Arabs who in 751 found paper makers among Chinese prisoners who were taken in a battle at Samarkand. The art spread through the Near East and to lands held by the Arabs, such as Sicily and Spain. From there and from Crusader contacts in the Levant, papermaking came to Europe and the stage was set for more systematic evaluation of the results of instruction.

The pioneers in the development of written tests were the Jesuits. Founded as a teaching order by a group of men who had been students together at the University of Paris, the Jesuits spent much time formalizing educational procedures on both secondary and higher levels. The "Ratio Studiorum", published in definitive form in 1599, contains explicit rules for the conduct of written examinations. These rules specify, for example, that students must be present at the stated occasion unless detained for weighty reasons; that they must be on time; that no one may speak after silence has been enjoined; that they should come supplied with all needful materials; that precautions be taken to prevent copying; that when the composition is turned in to the administrator, it cannot be returned;and that time limits be strictly enforced.

Oral examinations for the B.A. and M.A. were introduced at Oxford in 1636 as part of a long-needed reform. A new statute in 1800 dealt specifically with examinations and led to an honors program which emphasized high levels of proficiency in literature and mathematics. Written examinations were introduced at Cambridge and Oxford at about this time and in the year 1828 printed question papers were introduced. Written examinations were generally recognized to have been very influential in improving teaching and student achievement in the English universities and in time, with the introduction of civil service tests, written examinations were generally recognized in the Western world as appropriate for deciding who should be awarded degrees, who should be permitted to exercise a profession, such as law or teaching or medicine, and who should serve in a government post. Before leaving the subject of proficiency testing, we might mention that actually the ancient Chinese were the pioneers in this area. Civil service tests which they used for selecting members of their governing class, the mandarins, covered a period of history spanning approximately 3000 years with relatively few interruptions. Their testing, however, was not tied to instruction, since ancient China had no universities. The typical Chinese scholar seeking a government post characteristically spent long years in study of the classics preparing to demonstrate his abilities in open competitive examinations. Here there was no implication that skill resulted from the effects of a formal instructional program.

That proficiency and instruction were related would certainly have been admitted by the masters and doctors of medieval universities and by the mandarins of China. However, a problem that concerns education today is finding the specific sources of proficiency. If these sources can be identified and emphasized, the prospect becomes better for meeting the demands for higher degrees of proficiency and greater numbers of proficient individuals.

To carry out research in this area we need some way to differentiate between proficiency already acquired and the gain in proficiency that comes as a result of specific experiences. The problem is not an easy one because proficiencies of greatest interest require long periods for their development and manifest themselves in variable ways.

Among the specific questions of interest are these:

1. Can the effects of specific instruction be differentiated from the effects of other influences, such as growth, the social environment and self-instruction?
2. How can change in proficiency be measured?
3. How can change in proficiency be related to outside variables, such as measurable characteristics of the learner and of the instructor and variation in instructional methods?

For some time Contract ONR 816(14) and its predecessor Contract 816(02) have been concerned with just these issues. Certain methods have been proposed to attach these problems, methods which, I think, have some justification. Nevertheless in an area as complex as the appraisal of the learning of high order skills, we need continuous study to identify appropriate criteria both for over-all proficiency and for changes related to specific internal and external factors. I hope that the conference today and tomorrow will give us new insights in this area.

THE CRITERION PROBLEM IN LEARNING RESEARCH

Marion E. Bunch

I wish to emphasize that in my opinion it is of the utmost importance to use a performance criterion of different levels of mastery in research studies on learning. In this connection, there are a few basic factors that serve as the fundamental conditions with reference to which the criterion problem must be defined.

First, learning is an inference from performance. Learning is considered in terms of relationships between stimuli and responses. In general, the correct response is considered learned when it can be given regularly and dependably to the proper question or stimulus event. The principal point is that the increased probability that a stimulus will arouse a particular response represents learning to the extent that the greater probability is a function of training. The increased probability is, of course, neither the stimulus nor the response. The degree of learning refers to the closeness of the functional relationship that has developed between the stimulus and the response and is identified with the probability of the occurrence of the response under conditions of appropriate stimulation.

Now what properties or aspects of behavior can best serve as indices of the increasing probability that the stimulus will arouse the response as training progresses?

Presumably, well-learned responses exhibit, on the presentation of the relevant stimulus, behavior which involves the following aspects:

1. A high degree of accuracy in the performance of the response learned.

2. A significantly shorter reaction latency than occurred at the beginning of practice.

3. An increase in the rate or speed of the correct response.

4. An increase in the amplitude of the response.

5. Increased resistance to experimental extinction.

6. Increased resistance to retroactive inhibition from subsequent learning as compared to the amount occurring when learning has been stopped short of mastery.

7. Increased positive transfer to subsequent learning in similar situations.

8. A certain degree of generalization to similar stimulus events.

In determining the extent to which the degree of learning is a function of any experimental treatment which we might wish to use, we employ one or more of these aspects of behavior as an index of learning. These measures, however, are not exclusively a function of the degree of learning. Changes may occur in them which have nothing to do with the progress in learning. For example, rate of responding may on occasion be a more sensitive index of motivation than of degree of learning at the time.

The first five of the behavior changes that have been noted are the ones most often used as measures of learning. Some measures are highly specific to the material or responses being learned and some are employed in several varieties of learning situations.

In classical conditioning the primary measures employed are latency of response, percentage frequency, amplitude or amount of response, and resistance of the response to extinction.

In instrumental or trial and error learning, as in the acquisition of skills, solutions in complex problem solving, and understanding in cognitive learning, the performance measures usually recorded are correct responses, errors, and time scores in successive trials. Changes in these performance measures during learning are regarded as representing improvement and are used as indices of increasing probability of response. As one writer noted, a record of correct and incorrect responses during succeeding trials would seem "to approximate as closely as anything imaginable what is normally regarded as learning or the changes resulting from learning." (Bugelski, 1956)

As measures of learning there is good reason for regarding success-or-error scores as superior to latency of response, rate, or amplitude of response, in view of the fact that, first, the occurrence of errors in the initial performance of the act constitutes the best evidence that the subject is confronted with a problem for which he does not already have a ready-made response that is correct, and second, the correct performance without error after training provides the best evidence, or behavior measure, that the problem has been mastered. It is also true, generally, that as errors are eliminated during practice, later trials are completed in less time than was required in the early trials. However, if time scores are the only scores available in a learning experiment, interpretation is difficult and questionable.

The conventional graphical way of representing the progress in learning is in terms of a curve demonstrating improved performance, defined as a reduction in the number or percentage of errors, reduction in response time, or an increase in the number or percentage of correct responses, over the series of trials. These measures serve to carry the notion of increasing probability of response. We do not observe a probability as such; we describe the frequencies with which responses occur. Presumably the probability that a response will occur in a given situation may range continuously from zero to 100 per cent. Presumably also, the probability of the stimulus arousing the correct response increases during learning from an initial level of chance or near-zero probability, to the upper limit of learning afforded by the conditions of practice at which level the probability of the correct response occurring to the appropriate stimulus will be near perfect or 100 per cent.

In one view, emphasis is placed on an analysis of the stimulus complex. Any stimulus may be said to be composed of a large number of elements. The likelihood that a response will occur may be said to be directly proportional to the per cent of stimulus components attached to the response through previous learning. If none is attached, the stimulus complex will not arouse the response at all; if all components are attached through association, the response will occur inevitably on every trial.

In this view, the probability of the response follows from the proportion of the stimulus components that have become attached to it. Probability is assumed to be related to latency of response, to errors, and responses per minute, and these measures are often used to test aspects in the theory.

Percentage frequency of the correct act is especially important when the response may be expected by chance in some proportion. If the chance expectancy of the correct response is high, then a relatively high score is required as evidence that learning has occurred. For example, when an animal is faced with a right-left choice, there must be 15 consecutive correct choices within 3374 trials for the subject to be said to have learned the discrimination. If our experimental subject is a student and the task is that of memorizing a poem, practice may be terminated at some arbitrary point determined by the experimenter such as one perfect trial or three successive perfect trials. The problem is said to have been mastered when the arbitrary criterion has been met. At this stage in practice, the rate of rise in the negatively accelerated curve has gradually diminished to a degree that the curve has become virtually horizontal. Although the degree of learning may continue indefinitely, a more severe criterion than that just noted might be more concerned with the problem of making an automaton out of the person so that the performance of the well-learned act would not be influenced by such things as inattention and carelessness, than concerned with the task of promoting further learning. Hence, the arbitrary criterion of mastery is usually at the stage where the curve becomes asymptotic to a horizontal line.

In recent years, two of the most frequently used measures of learning in laboratory studies do not involve a performance criterion of mastery which a subject is expected to reach before practice is terminated. These two are: a fixed number of training trials which is the same for all subjects, and resistance to extinction. Both of these measures involve severe limitations and neither would appear to be a sensitive measure of learning.

Using number of reinforcements, or trials, as an index of degree of learning is really nothing more than a practical and useful procedure for either ranking, or equating, groups on the basis of the number of units of opportunity for the growth of the habit to be established. An independent measure of the actual increase in response probability produced under the different number of trials is seldom employed. The equal trials procedure ignores the individual variability in the degree of learning that results from the specified number of reinforcements and ignores also the very probable fact that the degree or amount of individual variability in learning is not the same for all stages of learning. A single reinforcement cannot be assumed to produce the same degree of learning in different subjects, and in all probability no two reinforcements in the same S during learning produce equivalent degrees of learning. Number of reinforcement trials instead of being a measure of learning would appear to be a procedure for producing a comparable but unknown amount of learning in similar groups of subjects. While the level of growth at any time in the life of a child is a function of the number of units of growth opportunity, a measure of growth is needed even to indicate that two comparable groups, e.g., 8 years of age, have attained comparable growth levels.

In many experiments the resistance of the learned response to extinction is the primary measure used by the experimenter. It is as if the learning itself is only something to get over and done with so that extinction can be measured. Learning is not continued until a performance criterion of mastery has been met, but for a fixed number of trials, and the degree of learning is inferred later on the basis of the number of trials required to extinguish the response. It is assumed that the greater the resistance to extinction, the greater the degree of learning that must have been reached during the fixed number of trials of practice.

There is considerable evidence in support of this position. For example, in Williams' experiment, on which Hull leaned rather heavily in suggesting number of reinforcements as a measure of degree of learning, the four groups which received respectively 5, 10, 30 and 90 reinforcements during acquisition, ranked exactly in that order in showing increasing resistance to extinction. However, the same result would have appeared if the second test had been one of transfer of training to a similar problem or a relearning of the original problem to provide a retention test. In other words, a transfer of training score or a retention score would have differentiated Williams' four groups according to the number of reinforcements which they had had during learning, as well as did the number of responses in extinction. Furthermore, some laboratory studies show that extinction is not the unlearning of the previously learned response but a procedure for reestablishing the degree of variability of behavior shown initially in the learning situation. In fact the degree of learning or strength of the habit may be as great in the experimental group following extinction of the response, as in a control group in which no extinction trials were permitted.

In fact, emphasizing the importance of a performance criterion of mastery and having the subjects continue learning until the criterion has been achieved, is really asking for an objective scale to constitute the vertical coordinate in order to show the extent to which degree of learning is a function of successive periods of practice, i.e., the units on the abscissa, under different experimental conditions.

One example may illustrate a number of the points involved in some measures noted above.

Curves A, B, and C, in Figure I represent respectively, the learning curves of a superior, average, and inferior groups of subjects in mastering a problem to the criterion of one perfect trial. As indicated, in this example, group A learned the problem in 10 trials, B in 20, and C required 30 trials to reach the same degree of learning.

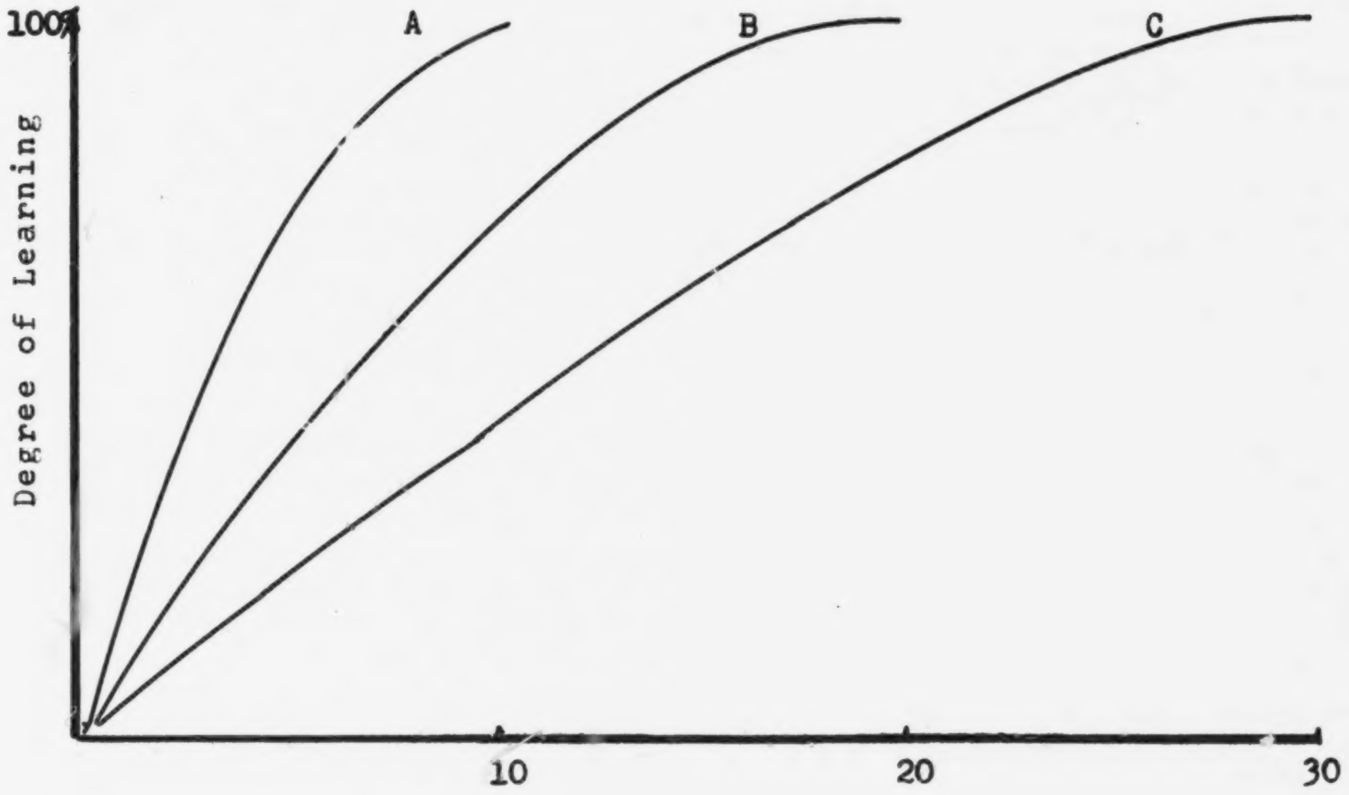


Figure I.

Let us examine a few questions which might be asked about these data:

1. If practice had been terminated for all groups at the end of the tenth trial, how would we expect the groups to compare with each other in a test of retention later, or in resistance to extinction later?

   On the basis of learning studies, I think the three groups would differ from each other and would rank in the order A, B, and C, with A showing the best retention score, or greatest resistance to extinction if that were the measure being used, group B would be next, and C would be in third place. In other words, the expectation is in accordance with the view that the behavior measures used in learning indicate that the groups differed in degree of initial learning and ranked A, B, and C.

2. Now if practice had been terminated in each group at the criterion of mastery, what would be the expected result?

For brevity, let us look at possible interpretations of one particular result, e.g., group C showing greatest resistance to extinction, with B and A following in that order.

a. Does this mean that resistance to extinction is a function of the number of trials in learning even when the degree of learning established during practice was the same according to the behavior criterion, or

b. Does it mean that the difference in resistance to extinction between the groups indicates that the degree of learning was not the same for the groups even though the behavior criterion indicated comparable level of mastery?

c. Does it mean that slow and fast learners differ in rate of extinction, thus accepting the behavior measures as indicating that the same degree of learning had been achieved during practice?

3. Now let us suppose that all three groups continue practice through trial 30, as shown in the figure. What would be the expected result as regards resistance to extinction?

   Again we would expect them to differ, and the rank order to be as follows: A, B, and C, thus reflecting the relative degrees of learning attained during practice. But without the performance measures in learning, with reference to the criterion of mastery, we would have no check on the relationship between degree of learning, on the one hand, including overlearning, and extinction measures on the other hand, and no way of checking on the assumption that trials to extinction are highly correlated with the degree of mastery attained during the corresponding number of trials in learning.

One of the major difficulties in measuring degree of learning is that, while the degree of learning at the time may be the principal factor determining reaction latency, probability of occurrence, magnitude of response, percentage of correct response, and resistance to extinction, may be influenced by other factors such as motivation and fatigue, and hence would be better regarded as measures of performance rather than of learning.

The view that the number of reinforcements or trials in learning is the superior measure and is perfectly correlated, or nearly so, with degree of learning is an inference and not based on a set of comparative correlation records. The same state of affairs exists as regards the inference that the number of extinction trials required to reach the performance criterion of extinction, constitutes such a sensitive measure of the degree of learning that behavior criteria in learning are unnecessary.

These are but a few of the possible questions and possible outcomes concerning which disagreement exists. Until there is greater clarification concerning the matter of acceptable criteria of degrees of learning during practice, the difficulty of testing alternative hypotheses will remain with us. And so, for example, we will continue to have some studies concluding that fast learners are superior in retention to slow learners, and other laboratory studies concluding they are equal in retention.

CRITERIA IN LEARNING RESEARCH AND SOME RELATED PROBLEMS OF DESIGN

Ellis B. Page

The announced topic of my paper is "Criteria in Classroom Research." This topic is clearly as broad as one could ask, as we have seen from some of the papers on criteria already presented this morning. If we paraphrase the topic into a question, such as "How can we organize the criteria of our research so that we extract from our experiment the optimum information?" then we may see that the problem of the criterion, properly viewed, is also the problem of the experimental design and the problem of the appropriate analysis.

In the last conference proceedings[1], I read a comment by Bryce Hudgins which shows an interesting analogy between educational experimentation and laboratory experimentation. He said, suppose you as a laboratory researcher were told that the rules had been tightened-that henceforth you were permitted to observe the rats only twice a week, for 46 minutes each time, and that you were not under any circumstances to manipulate the behavior of the animals. Such constraints truly surround much educational and training experimentation, and point up the desirability of making optimum use of available material.

The first two suggestions I will make are toward the more efficient analysis of the available data. And the last two suggestions have more to do with the conceptualization of the experimental unit. I shall not attempt to put all of these in one design package. This would be quite a complicated question, and the complexity would lack generalizability. That is, the combination of these ideas which would be appropriate in one experiment would not often be appropriate in another. Furthermore not all of these suggestions would normally be reasonably accessible in a single given problem of design. But as people from education or psychology have come to our bureau seeking help with design or analysis problems, these have struck one repeatedly as being undeservedly neglected ideas.

1) The first suggestion concerns the use of all appropriately available measurements. Clearly criteria will be multiple in many experiments. Consider the case of assessing the effect of a teaching or training procedure on a series of classrooms. In such a case very seldom is one measure adequate. In the typical school example, one may wish to look at the results of two or more standardized tests, and perhaps at a local evaluation of final essays and perhaps at some measures of attitude toward the subject. Campbell and Stanley have pointed out (in their chapter 5 of the N. L. Gage Handbook for Research in Teaching. Chicago: Rand McNally, 1963) that the day of contentment with the single criterion should be over.

---

[1] DuBois, Philip H. and Wientge, King M. (eds.) Strategies of Research on Learning in Educational Settings, Washington University, St. Louis, Missouri, February 1964.

What is typically done with various measures? They are tested in a succession of supposedly "unrelated" analyses. A doctoral candidate at a certain western university was happily describing (at his oral exam) his own thesis results which concerned analyses of 100 different measures, and his entire dissertation turned around the discussion of why five particular measures were "significant." The moment of truth occurred when the methodologist present quietly put his question: "Out of 100 measures," he asked, "how many would you expect to be significant at the 5% level-by chance alone?" The student hesitated, turned ashen in color, and gasped, "Five!"

Of course, there are various ways of handling multiple comparisons, but I would suggest considering the use of multiple discriminant analysis. Such analysis will combine the measures in the most efficient way to achieve significance, and will help answer the question, "In what ways do these treatments differ most importantly?"

Multiple discriminant analysis has been around for some years, but seldom is it employed in experimental, as opposed to psychometric situations. The gulf, lamented by Cronbach and others, between the experimentalist and the psychometrician still exists. To some extent it always will, for as the experimentalist learns more of psychometric point of view, the psychometrist learns a dozen new leads himself.

In consequence, we have such a major and thorough experimental book as Winer's Statistical Principles in Experimental Design (New York: McGraw-Hill, 1962) where there is no explicit indexing of multiple discriminant approaches. There are, however, standard computer programs for certain multiple discriminant designs.

2) A second area of consideration I would suggest concerns the relation of the statistical analysis to the fundamental interests of the scientist. This is that the experimenter consider especially for his main effects, the use of ordered hypotheses on the analysis of variance.

If the hypothesis is expressed by

$$H_o : M_1 = M_2 = \ldots\ldots\ldots\ldots = M_k$$

where M_i represents the mean for the i^{th} treatment, then the most general form of alternative may be written

$$H_1 : M_1 \neq M_2 \neq \ldots\ldots\ldots\ldots \neq M_k$$

H_1 is an expression of what may be called the "omnibus" F-test in the analysis of variance, and is the basis of nearly all existing work. Yet what is more often of scientific interest is a test of the null hypothesis against the ordered alternative

$$H_2 : M_1 > M_2 > \ldots\ldots\ldots\ldots > M_k$$

Not only is this of more exact interest, it is also much more powerful. Boersma, DeJonge, and Stellwagen proved this at length in a recent article in the *Psychological Review* (November 1964, pp. 505-613). "Significance" is vastly more easy to obtain with limited samples when the hypothesis is indeed true in the population (of events or persons) of interest.

What Boersma *et al* did was construct a good number of such populations in two-way layout by random generation in the computer, and to test samples of these for rejection of null hypothesis against 1) the usual omnibus F-test; 2) the usual ranking analog of this omnibus F-test, the Friedman two-way analysis of variance; and 3) by my own L-test, an ordered test for Friedman-type layout of multiple rankings (which I described in the *Journal of American Statistical Association*. March, 1963, pp 216-230). What Boersma found was that the L-test is vastly more powerful than either the Friedman test or the usual parametric F-test in the usual omnibus analysis of variance.

This power accounts for the considerable popularity the L-test has enjoyed since I published the paper. It is a way to "save" many dissertations and other researches from "nonsignificance"-a crucial practical career question for editors or advisors.

Of course, the L-test is not the only example of such ordered hypotheses, but the other possible approaches have not been set forth very clearly for psychological readers. If you would like to read more on the question, you will find an adequate beginning bibliography in my own article or in Boersma's. Truly, it is a severe waste of data *not* to use ordered hypotheses when they are appropriate. Yet again, no standard book on experimental design and analysis explicitly treats such ordered hypotheses.

3) A third consideration in optimum use of experimental material is that of blocking in of all appropriate available material. What I want to emphasize is the possibility of using incomplete block designs, and especially certain variants of this notion-using them respectably and with rigorous analysis. Consider the case which Stanley and others encountered in an experiment in which I subsequently became involved, where there was a limited amount of experimental material and there were a number of treatments which the experimenters (a group of teachers in a school system in Manitowoc, Wisconsin) considered important. The hypothetical available material is as follows:

HYPOTHETICAL AVAILABLE MATERIAL

Teacher	Number of Classrooms	Teacher	Number of Classrooms
A	5	H	2
B	3	I	2
C	5	J	2
D	2	K	2
E	5	L	1
F	4	M	1
G	4	N	1
		TOTAL	40

Different numbers of units are available in the potential sample. Until the last couple of years incomplete block designs were virtually unused in psychology or education. What is done is usually to use a one-teacher-per-cell design which would throw away two-thirds of this data, or to use some sort of complete-block design such as the well-known randomized block or "treatments x subjects" design familiarized by Lindquist or others.

Only in the last several years has the knowledge of incomplete block (IB) designs been given some psychological currency (very notable in Winer's 1962 text although Yates' contribution was in the 1930's). The IB design is simply a design in which the individual subject contributes less than a complete set of experimental treatments. Certain blocks are natural; if you are testing automobile tires, the natural k or block size is 4 although you may be interested in 8 or more makes of tires. If you are testing rubber heels, the natural or block size is, obviously, 2 with each walker serving as his own control, regardless of the heels investigated.

A bit of nomenclature: Balanced Incomplete Block designs (BIB) are designs in which every combination of treatments must be used the same number of times. In our classroom case, this would mean that, with a block size of 2, treatment 1 would need to be paired with every other treatment at least once.

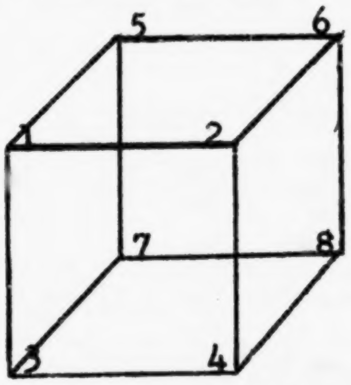
Partially Balanced Incomplete Block (PBIB) designs, on the other hand, are designs in which any pair of treatments occurs with one or two possible frequencies. In the present example, there is one rather elegant treatment of the material, assuming a restriction of block size K = 2.

ASSIGNMENT OF TEACHERS TO METHODS
IN A CUBIC LATTICE DESIGN

Teachers	**Methods**							
	1	2	3	4	5	6	7	8
A	X	X						
B			X	X				
C					X	X		
D							X	X
E	X		X					
F		X		X				
G					X		X	
H						X		X
I	X				X			
J		X				X		
K			X				X	
L				X				X

Note: Number of Teachers = 12, Methods = 8, Block Size = 2

The above layout may be produced from the following diagram in which each bar, connecting two treatments represents one pair of treatments assigned to some teacher.



This may be analyzed as a set of "quasi-factors" in which the systematic clusterings of unit contributions permit a partialling out of individual differences from the error terms. Again, both the BIB and the PBIB designs are considered in easily available texts today, though still little applied by the journeyman researcher in the behavioral sciences.

What is not covered is what I call the VKIB design (for Variable-K Incomplete Block). What is little recognized about IB designs is that, if one wishes, one may ignore the IB nature of the data and treat the data as consisting of complete blocks, without loss of rigor but only of sensitivity. Yates recognized this in 1939 (Annals of Eugenics, pp. 136-156) but the implications for our kinds of variable-K material have not been capitalized on at all.

You can lay out a complete - block 5 x 8 mating which (by no coincidence) nicely employs all the available data.

ONE ASSIGNMENT OF TEACHERS TO METHODS IN
A VARIABLE K INCOMPLETE DESIGN

Replications	**Methods**							
	1	2	3	4	5	6	7	8
1	A	B	A	A	B	A	B	A
2	D	C	C	C	E	C	C	D
3	E	E	F	F	E	F	E	F
4	H	G	I	G	G	I	H	G
5	J	K	K	M	L	J	N	O

Note: Each letter stands throughout the figure for a particular teacher; e.g., all examples of letter A represent different classes of teacher A.

About 25% in sensitivity is lost through treating it as complete-block, but much more is gained than that in additional experimental material. I do not want to over-simplify this analysis, but I do at least recommend it to your attention.

4) The last consideration is what I have been skirting in this paper so far. That is the appropriate unit of analysis in an educational or training experiment where there is group instruction over time.

Let no one try to catalog all the variables in any classroom or other instructional group: variables of student, teacher, curriculum, environment, variables measured and unmeasured, conceptualized or unimagined. They are countless, and their interactions are many times more numerous. Yet as scientists we must abstract and simplify, if we are to understand behavior. So we winnow and refine our measurements, paring away at this great mass, and seeking out rules which will lend our few discoveries some predictive power for future events.

The final portion of this paper is concerned with finding meaningful relationships among such classroom variables. First it shall be argued that the usual experimental classroom must be considered as if it were a single "subject." Then a way shall be presented to gain back something of the individual differences within such classrooms, still within a defensible statistical framework. This concerns, therefore, the treacherous problem of "independence" of observations specifically within the classroom, but those with a passion for larger generalizations will realize that the suggestions made here apply equally to the analysis of all intact human groups.

a) The Classroom as One "Individual"

Why must the classroom be considered as if it were only one "subject?" The question is not well understood, as judged by the literature, and will therefore receive some attention here. An illustration may be helpful: Suppose we have two classrooms, each having different treatment. Suppose that (in an unusual case) students may be considered randomly assigned to these classrooms. Suppose also (again usually not the case) that great pains are taken to make sure the instruction is as identical as possible, apart from the treatment. The classes are even given at the same time of day, in classrooms just across the hall. But an extraneous event (and this is usual!) is introduced: a gardener with a power lawn mower comes roaring by one classroom, causing some student to have hay fever, and others to lose track of the presentation in a crucial moment. This disturbance, although no part of the treatment, is assigned to one treatment group and not to the other. Then we have the case in Figure 1 where X, the lawn mower, happens to be in Treatment A_1.

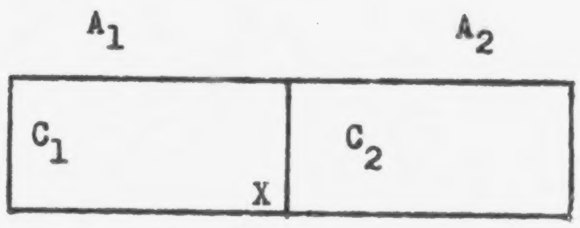


Figure 1

(C_1 refers to the ith classroom)

X could be any variable affecting the class as a whole; such as a particular teacher who depresses scores, a particular student whose noxious attitudes are contagious, or the particular timing of the class after lunch hour when students are sleepy. The point is that this "lawn-mower effect", as some of us have called it, is not ordinarily detectable through the usual procedures of experimental research or reporting.

There are no purely statistical safeguards against such a fallacy, although some of the previous literature might lead you to think there are. What is sometimes recommended with such errors (cf. Lindquist, 1952, passim) is to have at least two or three intact classrooms in each treatment. Then the comparison may be made,

$$(1) \quad F = \frac{\text{Mean Square (between groups within treatment)}}{\text{Mean Square (within groups within treatment)}}$$

If the result of such an F-test is not significant, according to this practice, one may safely pool degrees of freedom within the treatments and run the test

$$(2) \quad F = \frac{\text{Mean Square (between treatments)}}{\text{Mean Square (within treatments)}}$$

using new means averaged over all groups within treatments.

This preliminary test of Formula (1) may seem at a glance to protect us against incorrect assumptions of independence, but does it? Let us look at a slightly expanded design, seen in Figure 2.

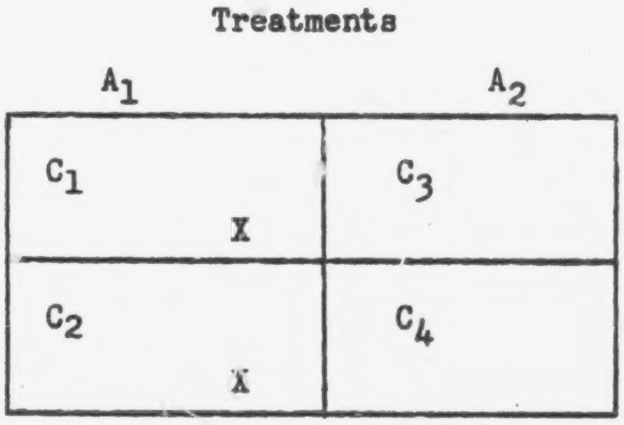


Figure 2

Here we see that, by no very great chance, this condition X, representing the lawn-mower effect, has occurred to two of the classes, and that these classes happen to fall within the same treatment group. It is clear that X will occur twice in Treatment A_1 about one time in four random assignments, and twice in Treatment A_2 about as frequently. Half of the time, then (rather than once in 100 trials or whatever alpha might be), the resultant depression of group score may severely bias the experimental conclusions. Yet in neither case would the F-test based on Formula (1) be successful in locating bias. On the other hand, if X were assigned as in Figure 3, occurring once under each treatment, then Formula (1) might detect "group influences," although in that case they would not necessarily harm generalizations from a pooled analysis.

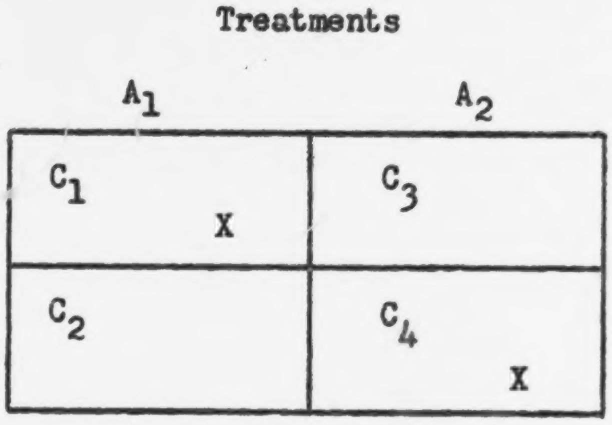


Figure 3

There appear to be three main sources of error in the assumption of independence within intact groups. One is the possibility of non-random assignment of subjects, similar to the Type S error so well known to us. The second is the sort of non-random events which may happen to a group to bias the results, such as the "lawn-mower effect" or other Type G errors. A third type of error, if indeed it is different from Type G, may be found in the interactions of communications of the students themselves within the classroom. We may call this Type C if we wish. Since these group and interaction influences are not detected by standard experimental procedures, they must ordinarily be assumed present. And the solution, therefore, is to move the randomization procedures, and the subsequent analysis of data, to the intact group as a whole.

Mistakes concerning such independence fill the classroom studies in the literature, and their results are rather easy to predict. They typically confound Type G and Type C errors with the experimental treatments, and consequently produce far too many "significant differences." These differences, then, are sometimes rendered unintelligible by the next study, which is apt to make the same mistake but with the random Type G or Type C errors differently assigned. The most defensible procedure for such studies, then, is to regard each such intact group as containing a single subject, and to conduct the analysis accordingly.

b) The Classroom as a "Repeated Measurement"

When the class is reduced to an "average," however, we seem to have reached analytic impoverishment. The substantive researcher is aghast at the suppression of all the intricate individual characteristics into a barren and general mean. He justifiably laments the loss of intelligence, sex, socio-economic scale, and a thousand other differences among the students in any given class. He longs again for the kind of "treatment-by-levels" design memorialized by Lindquist (Lindquist, E. F. Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin, 1953) and by others more recently. How can he regain this richness, as I used the term at the AERA meetings in February, within the classroom?

The answer may be rather simple, and can be expressed in a very compact rubric: Treat each such classroom, as we have said, as if it were a single subject. Then treat the interesting sub-categories within the classroom as if they represented repeated measurements of the same subject, made under different pseudo-conditions.

At first we seem to violate a sense of biological integrity in so considering "groups" to be "individuals" and then considering measurements of different sub-groups as if they were "different conditions" of the same group. Yet careful examination will show that such a way of regarding them is extremely useful, permits more or less rigorous treatment of the statistical aspects of such experimental material, and does indeed "recapture the richness within the classroom."

The principal virtue of the rubric should be evident: it permits us to enter the standard books of experimental design to the appropriate sections on repeated measurements (e.g., Winer, 1962, chapters 4, 7) and there find the correct layouts and analyses. It permits us to know also what is or is not testable. In so doing, it demonstrates the most important power of all good generalizations: permitting the transfer of a great deal of sophistication from one setting (i.e., repeated measures on an individual subject) to another setting (i.e., different sub-groups within the same intact experimental unit). While this generalization is easy to understand, it is not by any means trivial.

c) An Example

Let us suppose we have the situation pictured in Figure 4. Here treatments are labelled A_1 and A_2, each given to 3 different classrooms (which of course have been randomly assigned to treatments). These classrooms are represented by the rows C_1, C_2 . . . C_6. Such would be a rigorously "impoverished" design, in which each classroom is reduced to its mean score. Assume here that the columns represent four levels of intelligence of the students, for example, low, average, above average, and high, and assume that we have some interest in the interaction of treatment with intelligence, within such classroom settings.

I. Q.

Treatments			B_1 Low	B_2 Average	B_3 Above Average	B_4 High
	A_1	C_1				
		C_2				
		C_3				
	A_2	C_4				
		C_5				
		C_6				

Figure 4

In such a design, the summary analysis may proceed according to Table 1.

Table 1

Summary Table of Analysis

Source of Variation		df.	Error
Between classrooms		5	
A (treatments)	1		
Classrooms within groups (Error between)	4		Error between
Within classrooms		18	
B (Intelligence levels)	3		Error within
AB Interaction	3		Error within
B x Classrooms within groups (Error within)	12		
TOTAL		23	

This illustration is modeled after Winer (1962, pp. 303-312), where he was discussing repeated measures on individual subjects. This example simply demonstrates that one may, by this device, test for the interaction between the treatment and the intelligence level of students within such classrooms. Similarly, of course, other variables could constitute the pseudo-conditions represented by B here. Also, the designs may become quite complicated, though the possibility of various tests needs to be studied anew for each complication.

This is not the occasion to review the various problems concerning repeated measurements, for instance, whether the "usual" or "conservative" F-test should be performed here. These are fit subjects for extensive debate in the journal and textbook literature, and there they will be found. Perhaps a general caution should be invoked about error terms involving repeated measures, which do not have random between-classroom differences incorporated into the denominators of the F-ratios. In general, such repeated-measures errors are quite limited in generalizations across a population of subjects or classrooms. But here again, the restraints or equivocations are only as restrictive as the comparable ones about individual subjects.

It should be noted that some of the pitfalls of repeated measures are not a concern in such intact-classroom analysis as we have discussed. The problems of sequence or order effects of the treatments, so often encountered in the classic repeated-measures studies, are of no concern when different sub-groups within classrooms are really the various "conditions." On the other hand, one must keep in mind, when drawing inferences from experimental results, that generalizations should be, rigorously speaking, restricted to classroom conditions of the same sort, and from the same population, as were sampled for the experiment concerned. It would clearly be mistaken to slip back into the assumption that the same results would be achieved for the different intelligence levels, for example, if these levels were separated into homogeneous classrooms. Therefore, it is clear that the generalizations resulting from such analysis will often have predictive value, rather than promising immediate manipulative application to classes formed in new and special ways.

On the whole, however, a great deal of the interesting data within classrooms may be brought back without loss of respectability, into the light of statistical examination.

Conclusions

What I have tried to set before you this morning are some suggestions about increasing either the sensitivity or the rigor of experimentation in classrooms or other intact groups. We have many such studies in training, social psychology, merchandizing and similar areas of behavioral science.

Without both sensitivity and rigor, we are badly handicapped in our pursuit of correct criteria in learning research.

RESEARCH WITH "NORMAL" ADULT POPULATIONS: THE BROMWOODS STUDY

King M. Wientge

In analyzing the learning environments of day-division students and evening-division students, educators have commented on the marked differences in their social milieus. By this they have meant that full-time students in day-division programs have more opportunity to relate in meaningful ways with their peers and with faculty. It is assumed that learning is augmented and attitudes toward higher education in general and the specific institution in particular are favorably influenced by the day students' opportunities for peer group and faculty contacts. Evening division administrators, tending to emulate the day division offerings, have developed such evening counterparts as student honor societies, social groups, and have provided recreation activities to promote interaction among students and faculty.

In recent years there has been a movement toward the establishment of adult residence centers. Some of these, such as the Kellogg-supported centers are centrally located in parent campuses. Others are remote residential centers whose sponsors claim that more efficient and effective learning can occur when the adult enrollees are "out of reach" of their day to day job situations. The most recently established Kellogg Center is at Oxford University in England. There based on the assumption[1] "that a university is a band of scholars living as a community and working together in an intellectual pursuit," the Oxford version of a residential center is found at Rewley House.

One essential dimension that is lacking from the modern concept of residence centers is that of time. The early bands of scholars gathered in a university residence setting were cloistered for long periods of time as they exchanged information and shared experiences. One can logically raise the questions concerning the efficiency and effectiveness of the learning which occurs under the modern version of the residential experience. Many workshops, conferences, institutes or short courses are conducted in the course of a year at the typical adult residential center and the annual reports of these institutions list the number of different adult groups that have been served and the gross number of adults that have "been through" the center in the course of the year.

It seems timely that researchers in adult education look systematically at the influence of the residential setting upon adult learning. Sicuro[2] in a study of a sample of the freshman population at Kent State compares the off-campus center student and the central campus student.

---

1. Jessup, Frank, Secretary, Delegacy for Extra-Mural Studies, University of Oxford, Continuing Education Report, Number Four, University of Chicago, 1965.
2. Sicuro, Natale A., A Comparison of Academic Aptitudes, Certain Values, and Personal Background Characteristics of Students in Off-Campus Centers and on Central Campus of the Same University. Journal of Educational Research, Vol. 58, No. 5, January 1965.

He suggests two alternatives to be solved by future research. "Further, if it be ascertained through other research that academic center students are little interested in the social life of a campus, and should follow up studies substantiate that they can succeed in college without partaking of those extra-curricular activities which presumably contribute to the social development of college students, it may be questioned whether colleges need to provide the experiences. Such a possibility even raises doubts about the value of residence on the campus as a requisite for scholarship.

On the other hand, if further research provides a positive correlation between college success and participation in socializing experiences which a campus ordinarily affords, there are implications here for expanding the extra-curricular provisions at the centers in order to compensate for the lack of activities in the secondary school background of most center students."

Sicuro's subjects were drawn from freshmen attending the central campus at Kent State University and freshmen attending the eleven off-campus centers under the jurisdiction of Kent State. The similarities of the latter group of students to adult students are worth noting. They were economically unable to attend full time college programs, many were married and employed and their courses were scheduled after working hours.

The Bromwoods Study was undertaken to measure the impact of socialization factors on adult student achievement and adult student attitudes in a credit class of beginning psychology. It provides information about the amount of learning which occurs under standard teaching conditions for an evening division credit class in general psychology as compared with a similar class taught under experimental conditions, which involved two weekends at Bromwoods, the residential center of Washington University. It was hypothesized that the closer, more intimate student-teacher and student-student relationships, which would be developed by two weekends together in a residential setting would result in significant increases in learning and significant positive changes in attitudes for the experimental class as compared with the control class.

RESULTS

The objective results obtained did not support either hypothesis of significantly greater improvement in attitudes or significantly greater learning by the experimental group. The amount of psychological material learned did not significantly differ for the two groups. The measures used were 100 item objective tests of general psychology. They were administered to the experimental and control groups at the beginning and at the end of the course. No significant differences were obtained.

The analysis of the attitude scores showed a marked consistency in measured attitudes for the experimental and control groups before and after exposure to treatments. There were no significant differences present in measured attitude toward Washington University, University

College, college instructors, evening students, psychology classes, or evening school classes between experimental and control groups at the beginning or at the end of the study. These results are puzzling in their consistency. No "halo" effect is apparent for the experimental group. Does this mean that "halo effect," in fact, was not present or that the measuring instrument was inadequate, i.e., not sensitive enough to detect any changes. Finally, of course, there is the possibility that the objective attitude scores were entirely correct and that significant changes in attitude did not occur for either experimental or control groups. The attitude measures used in the study were based on concepts of the semantic differential and permitted a 1 to 9 rating of the attitude being measured. It is apparent that instruments used to measure changes in adult learning in classroom situations should have some sort of a priori validity established.

An evaluation questionnaire was administered to the experimental class at the end of the course. Of the twenty questions on the questionnaire several pertained to attitude change. On these questions there was substantial agreement, (11 to 2) that the participants had enjoyed the experimental course and would relish the same arrangements in another course. The same agreement existed in their statements pertaining to knowing their instructors and fellow students better.

The disparity which exists between results obtained with the questionnaire and the attitude scales suggests further exploratory studies of adult classroom populations in action as well as the development of more effective measuring instruments for sensing change.

A further consideration concerns the continued application of criterion measurements over time. No difference at the time of examination cannot be construed as an indication that no difference in retention will exist. Is the influence of the more favorable residential environment such that significantly more material is retained over time?

In sum, these comments describe briefly an attempt to introduce adult students into a favorable social milieu which, it was hypothesized, would produce significant changes in learning and attitudes. The fact that neither occurred was examined from the standpoint of the sensitivity and validity of the instruments used and of the possible relationship between time and criterion measurement.

Brownell[3] presents an insightful discussion of the evaluation of learning under different conditions and kinds of instruction. His discussion covers the complexity of evaluative research; the need for judgment by the experimenter; and the "common sense" evaluation of findings of statistical significance. In the latter sense criterion measures need to be developed that not only satisfy experimental rigor but also have practical implications for eventual practical applications to ongoing learning situations.

---

[3] Brownell, William A., The Evaluation of Learning Under Differing Systems of Instruction, E. L. Thorndike Address, Educational Psychologist, Vol. 3, No. 1, November 1965.

ABILITY TO LEARN OPERATIONAL EQUIPMENT AND SYSTEMS AS A CRITERION IN TRAINING RESEARCH

G. Douglas Mayo

Over the years the development of a practical criterion of on-the-job performance in training research has proved as elusive to the researcher as it has appeared obvious to the uninitiated. The criterion of school grades which was considered more or less an interim criterion as far back as World War II continues to hold the position of the most widely used criterion in military training research today, just as it did twenty years ago. This is not to say that no effort has been expended in criterion research, nor is it to say that no progress has been made. It does emphasize however that the problem is still with us and that it is a formidable one.

There are a number of reasons why on-the-job performance data have not proved entirely satisfactory. And a number of these are by no means solved by the approach described in the present paper. However, a criterion measure is proposed here which avoids a number of difficulties inherent in many on-the-job performance measures. The criterion proposed is the ability to quickly and adequately learn operational equipment and systems following general, theoretical training. In theory, at least, mastery of operational equipment which will be used in the immediate work situation ensures technical competence to a substantial degree.

Paradoxically, a change in training concept, dictated by the introduction of increasingly complex weapons systems, made the new criterion possible. No longer could a recruit be expected to learn to maintain operational equipment following general, theoretical training applicable to his aviation occupational speciality, simply by informal training in the equipment on the job. It was now necessary to provide formal training in the exact equipment which the man would be expected to maintain.

The concept of training which has emerged in naval aviation retains the economy inherent in mass production for general, theoretical training. Following this training personnel are assigned to an operating squadron but are trained on the specific equipment used by this squadron before reporting to it. This training is usually conducted at the naval air station at which the operating squadron is stationed when ashore, although the squadron may be deployed aboard a ship while personnel ordered to it are being trained at the naval air station. The organization doing the training has operational aircraft of the same type as the squadron for which they are training, but most of the training of maintenance personnel is conducted on training panels which are provided by the weapon system manufacturer, complete with all training materials required.

Most of the training equipment is essentially the same as that in the weapon system or aircraft but has been displayed in such a way that it can be more readily learned than would be the case in the actual aircraft.

Typical training equipment would consist of approximately 8 to 10 major instructional units, each unit consisting of 1 or more training panels and supporting materials. The costs of this equipment ordinarily would be between 4 and 5 million dollars. The instructional situation in which this training is conducted is typically a school type building of approximately 12,000 square feet of floor space. The atmosphere of training is that of formalized training in a trade school environment.

As stated a moment ago the present concept in naval aviation maintenance training is to provide newly enlisted personnel with the minimum amount of general, theoretical training which will permit them to learn the specific equipment on which they will be working during their first enlistment quickly and effectively. Since this is the objective of the general, theoretical training, an appropriate criterion of its effectiveness would appear to be a measure of the degree to which personnel completing this training are able to assimilate training on the specific equipment on which they will be working in operating squadrons.

When conceptualized in this manner, certain advantages are noted.

1. A substantial period of behavior observation is provided, longer than ordinarily would be possible in a performance testing situation. Most courses in which the observations are made range between two days and two weeks in length.

2. The behavior observed is directly related to the objective of the general, theoretical training and to the work the man will be doing on the job.

3. The observation of behavior occurs before further training and experience have an opportunity to have a differential effect upon the various individuals in the group.

4. The equipment and facilities associated with the observations of behavior are so costly that one could hardly hope to acquire them exclusively for training research purposes.

5. Since the situation in which the criterion data are collected is a learning situation, it is reasonable to expect that measures taken here, in addition to measuring performance at a given point in time, may also relate to ability to continue to learn operational equipment as the need arises.

Certain disadvantages, or remaining problems, might also be mentioned, for example:

1. The problems of measurement per se are still very definitely with us.

2. The usual difficulties associated with collecting data in an ongoing situation are still prevalent.

3. In a sense the researcher has as many different criteria as there are weapons system trainers or even courses, rather than having one convenient universally applicable criterion.

4. The number of subjects assigned to the individual courses tends to be small, thereby involving the problem of time required for collection of data, and problems in combining data pertaining to different courses.

5. Finally, the criterion does not have the characteristic of inclusiveness to the extent that would be desirable since even if the problems associated with combining the measures taken on all of the weapons systems trainers were solved there would still be some duties assigned to naval aviation maintenance personnel that would not be included. This results from the fact there are several of the older aircraft for which there is no weapons system trainer.

Accordingly, as noted earlier, the criterion proposed is not envisioned as a panacea, even within the limited context of naval aviation maintenance training in which the conditions are probably more conducive to its use than elsewhere.

We have completed one initial study in which the proposed criterion was used. The study reflects some of the problems just mentioned, and clearly is not intended as a model. It does, however, provide experience in the application of the proposed criterion and, hopefully, points to ways in which its application may be improved. First, I would like to describe the study and then mention some refinements or changes in the design that might be expected to improve the application of the criterion in future studies.

The initial study involved 231 graduates of Navy avionics courses. Approximately one-half of the subjects were given a course in avionics fundamentals only. Members of the other group had an additional two to three months of theoretical training in a more specific area of avionics. The operational question being asked was whether or not the additional training was necessary.

The criterion measure devised consisted of a form on which was listed 24 items of knowledge or skill which were taught in the second course but not in the first or fundamentals course. To the left of these 24 items was a column in which the instructor in the specific equipment or systems courses (that is, the criterion course) indicated whether or not each item of knowledge or skill listed was pertinent to the course in which the criterion data were being collected. To the right of the list of knowledge or skill items was a four-point scale with the heading "Effect of Man's Prior Knowledge of Item upon His Ability to Learn Material Contained in the Course." The four categories on the scale had the following descriptive phrases, as shown in Table 1:

1. Prevented adequate learning of the material.

2. Made the course more difficult or less effective.

3. Permitted learning the material in a satisfactory manner.

4. Greatly facilitated learning the material.

At the bottom of the form was an item involving evaluation of the overall effect of the man's prior knowledge of the pertinent items upon his mastery of the course, expressed in terms of the four categories which were just mentioned. This summary comparison of the ability of the graduates of the two avionics courses to assimilate training on specific equipment and systems is shown in Table 1. These figures pertain to the first specific equipment course taken by the members of the two groups. Although most of the students took more than one specific equipment course, the first one taken was considered to provide the best opportunity for evaluating the preceding general, theoretical training, since any course taken after the first course would be influenced by the material learned in this course.

Table 1

Summary Comparison of Ability of Graduates of Two Avionics Courses to Assimilate Training on Specific Equipment and Systems (First Specific Equipment Course Taken)

Course Taken	Over-all Adequacy of Previous Training								
	Prevented Adequate Learning of Material		Made Course More Difficult or Less Effective		Permitted Learning Material in Satisfactory Manner		Greatly Facilitated Learning Material		Total
	No.	%	No.	%	No.	%	No.	%	
Less Comprehensive	4	3	40	32	81	64	1	1	126
More Comprehensive	2	2	12	11	79	75	12	12	105

If the two categories to the right are considered to indicate satisfactory performance and the two categories to the left, less than satisfactory performance, we find approximately 35% of the men who took the less comprehensive general course performing poorly as opposed to 13% of those who completed the more comprehensive course. Applying the chi square test to the figures shown in Table 1, we find a chi square value of 14.19 which is significant well beyond the .01 level.

As indicated a moment ago, most of the students took more than one course on specific equipments or systems. Table 2 shows a summary comparison for the two groups on the basis of all such courses taken. As already indicated, the figures are not considered to be as good for criterion purposes as the figures pertaining to the first specific equipment course taken. Rather, they are shown as a matter of possible interest and for any additional information they may provide. Since more than one observation was obtained on the same individual, the chi square test is inappropriate and was not applied.

Table 2

Summary Comparison of Ability of Graduates of Two Avionics Courses to Assimilate Training on Specific Equipment and Systems (All Specific Equipment Courses Taken)

Course Taken	Over-all Adequacy of Previous Training								
	Prevented Adequate Learning of Material		Made Course More Difficult or Less Effective		Permitted Learning Material in Satisfactory Manner		Greatly Facilitated Learning Material		Total
	No.	%	No.	%	No.	%	No.	%	
Less Comprehensive	10	4	69	30	150	64	5	2	234
More Comprehensive	6	4	22	14	114	74	13	8	155

The primary purpose in this paper was to point to the criterion aspect of the study, and the figures given in Tables 1 and 2 are simply presented as examples, rather than a firm answer to the operational question. It should be mentioned, however, that while the groups assigned to less comprehensive training and more comprehensive training were so assigned in a manner which ordinarily would have made them comparable, a check on the aptitude of the two groups indicated a non-significant difference favoring the group that had the more comprehensive training. Somewhat more serious perhaps were the results of the comparison of the two groups on the basis of grades made by the two groups in the less comprehensive course, in which the group which was given further training was higher by a difference which was significant at the .05 level.

By way of summary concerning the initial study, it was largely an exploration into the possibility of using ability to learn operational equipment and systems as a criterion in training research. After having had some limited experience with it, in what ways can it be improved?

As a point of departure let us review briefly the disadvantages, or remaining problems pertaining to the use of the proposed criterion, that we listed earlier. They were as follows:

1. Problems of measurement, as such;

2. The usual difficulties associated with collecting data in an ongoing situation, further complicated here by the large number of separate courses;

3. The absence of unity in the criterion, also resulting, in part, from the large number of weapons system trainers and courses;

4. The relatively small number of subjects assigned to the individual courses; and

5. The lack of inclusiveness of the criterion, since some maintenance personnel are assigned to older aircraft for which there is no weapons system trainer.

If we formulated a completely adequate answer to these problems, we would have achieved something here at Bromwoods which has defied training research personnel for quite some time. But perhaps it would not be too ambitious to hope that we can chip away a bit at the problems. It may be noted at the outset that most of the problems relate in one way or another to the uncontrolled conditions which begin as soon as the students leave the schools conducting general, theoretical training. These include a time differential ranging from two weeks to more than a year in reporting for training on specialized equipment and systems, different difficulty levels of the courses taken, different treatment (in the broad sense of the term) or experiences of the subjects prior to assignment to training on the specific equipment and weapons systems, and informal selection of personnel for the specialized courses by squadron personnel. It probably is incorrect to assume that these and other uncontrolled factors will affect the members of two or more treatment groups in a random fashion. Such may be the case in one experiment, but not in the next. This is not a very adequate framework in which to conduct training research.

The following is proposed as a means of gaining some degree of control over the situation. The first step is to select one, or a small number, of courses on specific equipments or systems that are representative of the equipments to which first tour personnel in the occupational speciality in question are assigned. The course, or courses, should be representive in terms of difficulty, length, and content. The selection of the course(s) might be accomplished by means of a nominating procedure, to reduce the approximately 500 courses to a manageable number, and then by means of a panel of judges who are knowledgeable in the overall specialized equipment training area. The next step would be to assign members of groups receiving different treatments in the general, theoretical schools to the selected equipment course(s) either randomly, or as members of groups matched on the basis of a pertinent variable. The third step would be to devise the best possible measures of the performance of personnel in the specialized course(s) selected. This step becomes practicable in the case of a single course or a small number of courses, whereas it was quite impracticable in the case of some 500 course to which personnel might be sent.

It is recognized that the above proposal departs somewhat from the operational situation to which it is desired to generalize. It is thought, however, that this departure is not a very serious matter, and that its undesirable aspects are outweighed by the desirability of gaining control of the conditions under which the criterion data are collected.

This, in essence, is our current thinking on a refinement in the design of the initial study, which may move the basic concept one step nearer to adequacy as a criterion in ongoing training research.

THE CURVILINEAR RELATIONSHIP BETWEEN KNOWLEDGE AND TEST PERFORMANCE: FINAL EXAMINATION AS THE BEST INDICANT OF LEARNING

Ronald P. Carver

I. Introduction and Preview

The final exam has long been criticized as not being a good measure of classroom learning. Many psychologists have pointed out that there are individual differences at the beginning of a course which correlate positively with the final exam. Therefore researchers many times deduce that any measure of classroom learning should take initial individual differences into account when one cannot assume zero or equal initial differences. It will be my task today to suggest a third alternative. That is, although there may be high correlations between initial and final test performance, I will suggest that the highest indicant of learning is still the final exam itself. Stated differently, I feel that one need not necessarily assume either equal or zero initial difference in order to use the final exam as the best indicant of learning. I shall not only present this third alternative, but I shall present logical considerations which have convinced me that it is the alternative which best fits most classroom learning situations.

You may have noticed that I used the term "indicant" in my title. I am using the term exactly as S. S. Stevens does in his Handbook of Experimental Psychology (1951) pp. 47, 48. I consider this distinction between measure and indicant to be so important that I would like to quote what Dr. Stevens has written. I feel that too often this distinction has not been made in our quest for the best criterion of classroom learning.

> "Although psychologists devote much of their enthusiasm to the measurement of the psychological dimensions of people, they squander more of it in an effort to assess the various aspects of behavior by means of what we may call indicants. These are effects or correlates related to psychological dimensions by unknown laws. This process is inevitable in the present stage of our progress, and it is not to be counted a blemish. We know about psychological phenomena only through effects, and the measuring of the effects themselves is a first trudge on the road to understanding.
>
> "The end of the trail is measurement, which we reach when we solve the relation between our fortuitous indicants and the proper dimensions of the thing in question.
>
> "In the meantime we take hold of our problems by whatever handles nature provides. We count the number of pellets hoarded by a rat in order to assess its hoarding drive. We count the number of trials required for a man to learn a task, and use this number as an index of his ability. We measure changes in the resistance of the skin and

call it an indicant of emotion. In short, we are far more frequently engaged, as the following chapters will demonstrate, in the measurement of indicants than we are in devising scales for the direct assessment of physiological and psychological phenomena, or of 'intervening variables', as they are sometimes called.

"Occasionally the measurement of an indicant is sufficient for the task at hand; e.g., when we gauge a worker's ability by his productivity we may be interested in no more than the relation between his production and that of his neighbor. But more often we would like to measure his ability, intelligence, drive, emotion, hunger, etc., on a scale of the attribute in question rather than by effects that bear a dubious relation to it.

"The difference, then, between an indicant and a measure is just this: the indicant is a presumed effect or correlate bearing an unknown (but usually monotonic) relation to some underlying phenomenon, whereas a measure is a scaled value of the phenomenon itself. Indicants have the advantage of convenience. Measures have the advantage of validity. We aspire to measures, but we are often forced to settle for less.

"This distinction between measures and indicants disappears, of course, as soon as we learn the quantitative relation between the indicant and the object of our interest, for then the indicant can be calibrated and used to measure the phenomenon at issue. We measure electric current by means of a calibrated indicant composed of a coil of wire suspended by a spring in a magnetic field. We measure psychological pitch with a frequency meter after we have established a scale relating pitch in mels to frequency in cycles per second. The more mature a science, the more it uses calibrated indicants."

Today I will suggest the form of the relationship between test performance, the indicant, and learning, the variable which we are attempting to measure. Again from the title of this paper comes the suggestion that the form of this relationship is curvilinear, not the linear relationship as is usually implicitly assumed by researchers.

II. Discrepancies Between Logical Expectations and Previous Experimental Results

Most researchers have found low, zero, and negative correlations between crude gain (final score minus initial score) in learning some task and any other variable. Also, it has been pointed out by Manning and DuBois (1963), among others, that crude gain ordinarily correlates negatively with initial status. This does not seem logical since in classroom learning it often appears that bright students who have more knowledge at the beginning of a course also learn more so the correlation should logically not always be negative.

Of course, in many learning tasks all subjects approach the limits of the task. That is, they all approach mastery, and therefore, it is logical that those students who start with less, learn more. However, in classroom learning it is doubtful that all students approach total mastery of all the material which has been required for the course.

DuBois and Manning, in a number of papers, have presented a measure called "residual gain" (final score minus that final score predicted from initial score by a regression equation) which eliminates many of the disadvantages of crude gain. One of its major properties is that, by definition of construction, it always correlates zero with initial status.

After much thought and consideration, I decided that residual gain was a much better indicant of learning than crude gain. However, I felt that the most desirable measure of learning would be one which was relatively independent of both initial and final performance, independent in the sense that it could have a varying relationship with initial and final performance. That is, the best indicant of learning would be one which did not arbitrarily correlate zero with initial performance but would allow one to investigate empirically the relationship between initial performance and learning. Such was the background for curvilinear relationship which I will present under Section III.

III. Presentation of Model

Figure 1 presents the form of the relationship between the indicant, test performance, and the variable knowledge. I have arbitrarily labeled the variable which we desire to measure "Knowledge." At this point knowledge is a hypothetical construct or an intervening variable. Also, I will arbitrarily define classroom learning as crude gain in knowledge (final knowledge minus initial knowledge). Again, using Stevens' distinction: knowledge is the variable or dimension which we are attempting to measure, while test performance is the variable which is an indicant of knowledge. The graph in Figure 1 presents the suggested curvilinear relationship between test performance and knowledge.

Notice from the figure that a gain in test score from 65 to 70, a crude gain of 5, represents a greater gain in knowledge than the greater gain in test performance from 45 to 55 (82 and 37 units, respectively). This is the type of relationship of which Lord speaks in the Journal of Education and Psychological Measurement, (1958). I would like to quote Lord on this point:

> "...the gain of the good students do tend to be numerically less than those of the poor students. However, who is to say but that a gain from an initial true score of 65 to a final true score of 70 may not in every important sense be "greater" than the numerically larger gain from 45 to 55? The former gain for example, may represent more hours of study or more effort on the part of the teacher or perhaps a more important insight than the latter, numerically larger, gain."

The model in Figure 1 encompasses these thoughts expressed by Lord.

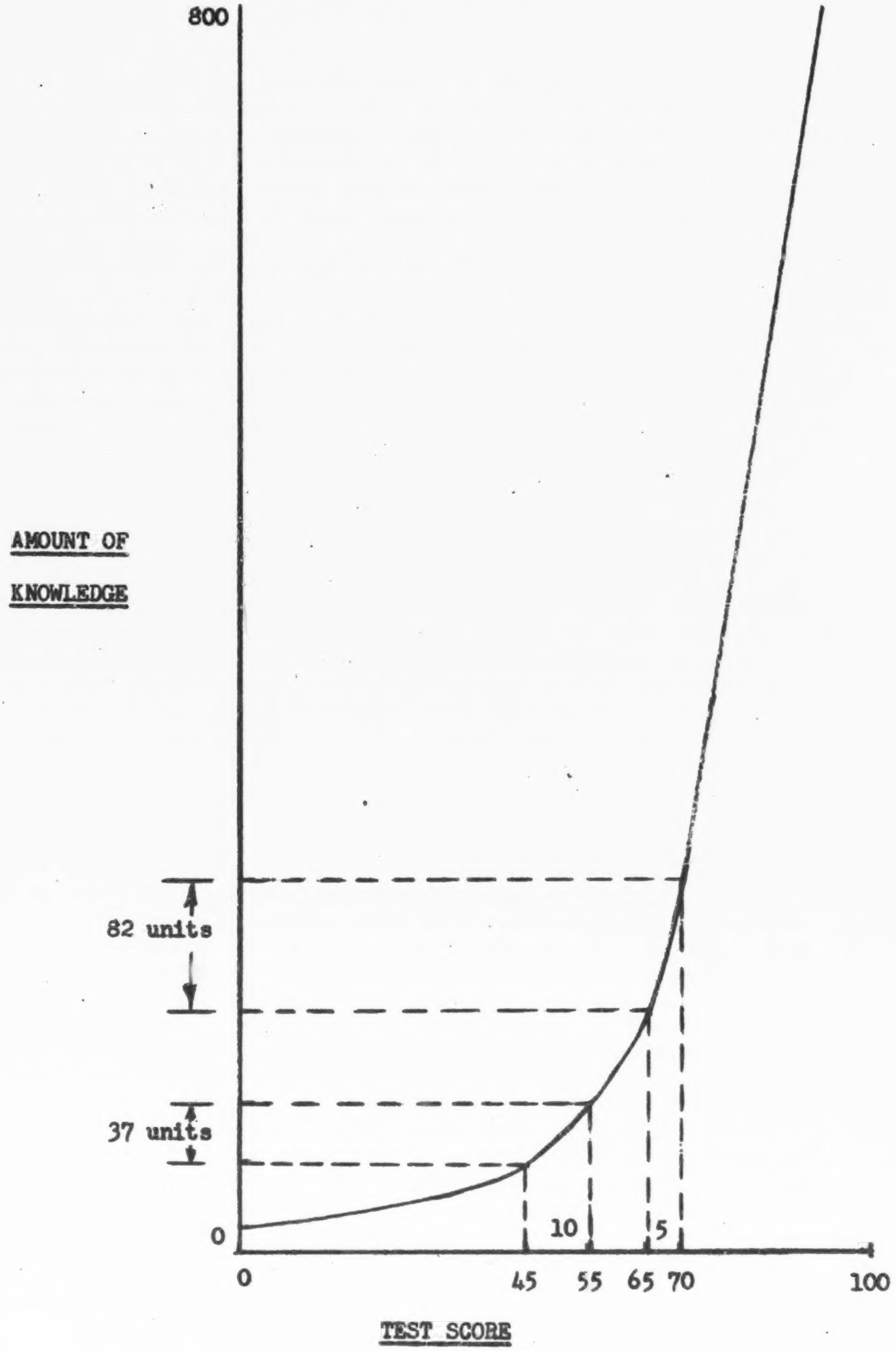


Figure 1. Amount of Knowledge as a Function of Test Score

Figure 2 presents the model in terms of initial and final exams on classroom learning. Notice on the figure that the variance of the initial test scores is approximately equal to the variance of the final test scores. However, the variance of the final knowledge is much, much greater than the variance of the initial knowledge. I feel that this high ratio of the variance of final knowledge to initial knowledge is the crux of the model. For when the ratio is as high as it is in this figure, it means that subtracting initial knowledge from final knowledge will not appreciably change the relative ranking of final knowledge. Stated differently, when the ratio of the variance of final knowledge to initial knowledge is high, the correlation between learning and final knowledge is near perfect and this correlation approaches unity as the ratio approaches infinity, regardless of the correlation between initial knowledge and final knowledge.

$$r_{LF} \longrightarrow 1.00 \quad \text{when} \quad \frac{F}{I} \longrightarrow \infty$$

Later, I will present an empirical example of actual test scores made to fit this model but for the present I want to point out the final deduction to be made from the model. Since the learning variable will correlate almost perfectly with final knowledge it follows that final test scores will also correlate almost perfectly with learning since from the graph it can be seen that final test performance correlates almost perfectly with final knowledge. Thus, at least in the case of this model, final exam seems to be the best indicant or highest correlate of learning.

A good fit of the model would also explain how all the past empirical research which used some combination of initial and final scores could yield results which are contrary to logical expectations. The answer might be that they assumed a linear relationship between their indicants and knowledge when the relationship was closer to being curvilinear.

Before going on to the next section I would like to suggest that if the model does provide a good fit to the classroom learning situation then it could also possibly explain the success that "residual gain" has enjoyed up to date. For if final exam is the best indicant of learning then residual gain could not be far behind since residual gain in most all empirical instances correlates very highly with the final scores, $r_{(F.I)F} = \sqrt{1 - r^2_{IF}}$

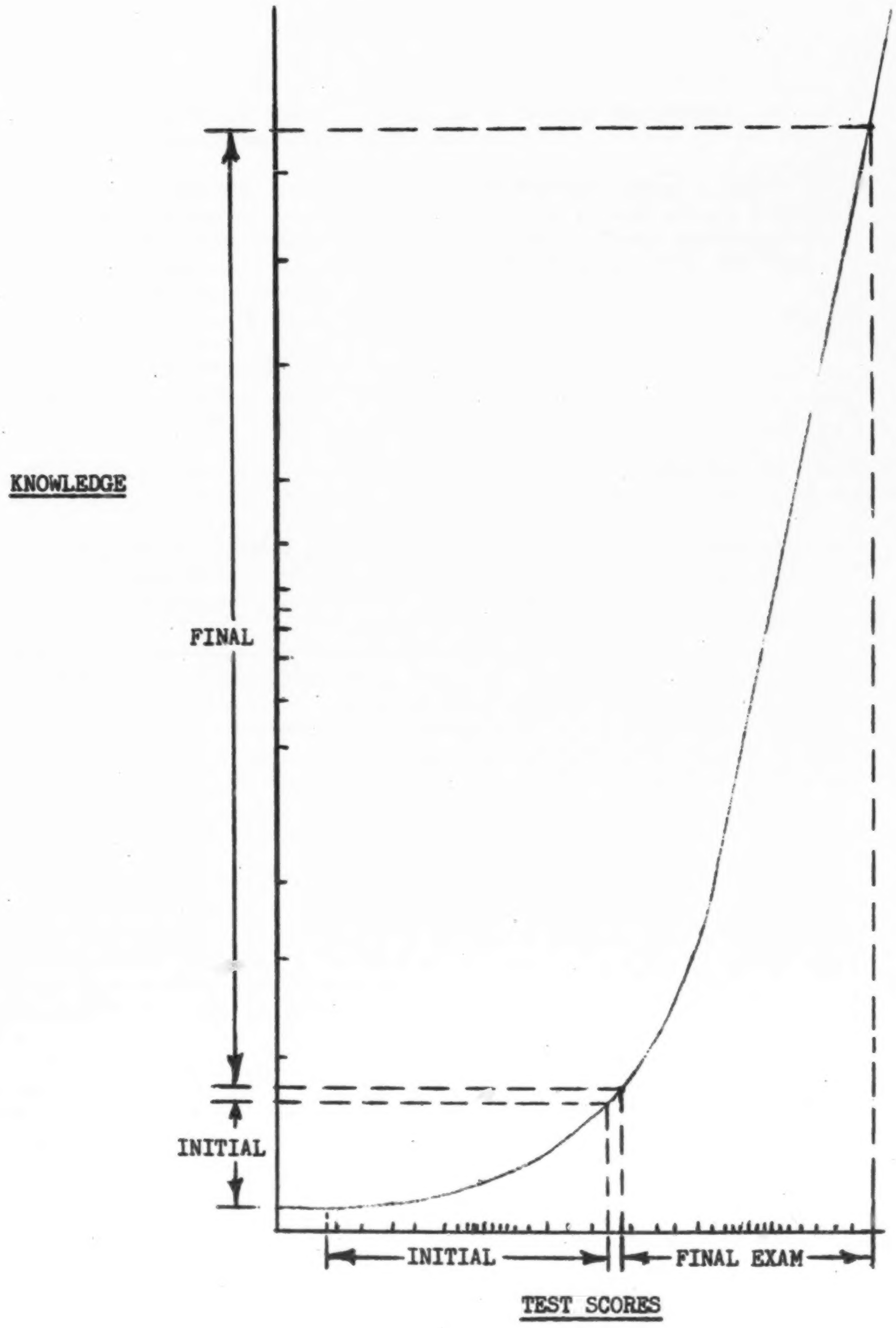


Figure 2. Knowledge as a Function of Test Scores with Initial and Final Distinctions

IV. Mechanisms Which Would Contribute to Making the Model Fit the Classroom Learning Situation

I have pinpointed four factors or mechanisms which I think would contribute to a curvilinear relationship. As I said before I feel that the crux of the model is the high ratio of final knowledge to initial knowledge while at the same time this same ratio for test scores is much, much lower.

F = final knowledge
I = initial knowledge
f = final test score
i = initial test score

Therefore the four factors which I will discuss all concern this differential ratio of variances when comparing knowledge with test performance.

Also, prior to the following discussion it is necessary that I explain in more detail what I mean by knowledge. That is, now I am going to give you what I think is a good conceptual operational definition of knowledge. When I speak of knowledge I want to refer to chunks of information which are much smaller than that which is often indicated by one test question. I am not prepared to specify exactly what I mean by knowledge, but if an item writer writes a question which encompasses a page or more of a written text he has not yet reached the units which I am thinking of. I want to consider the chunks of information which are contained in each sentence or at most in each paragraph. In a general psychology textbook I can conceive of about 1500 units of knowledge when considering only paragraphs, while if one considers each sentence then there would be about 15,000 units of knowledge. The word "bit" of information would seem to fit but it has a precise meaning in information theory which is very different from the unit of knowledge which I am thinking of. The term "chunk" as it is used in information theory is very similar to the way in which I am using the term. As stated previously, the following four mechanisms are considered factors contributing to the curvilinear relationship.

A. No Question. If one constructs a comprehensive exam which covers Ruch's general psychology textbook Psychology and Life and then administers it at the beginning of the course and then at the end of the course he will find that he has not written questions on many chunks of information. Since there are no questions on many chunks of information, there is a large source of variance on the final knowledge which is not indicated by the test scores. For example, some students will know many details plus the conclusion of a particular experiment, some students will only know the conclusions, while some students will skip over the experiment entirely. This represents a considerable amount of variance in final knowledge which is not represented in either the initial or final exam and which does not exist in initial knowledge, since none of the students know anything about the experiment prior to taking the course. Multiply this situation by the many areas which are not covered on the comprehensive exam and one can begin to appreciate how the ratio of final knowledge to initial knowledge could be much higher than the ratio of final exam scores to initial exam scores.

B. Question Threshold. Consider a test question which has been written on a topic which includes two pages of written text. Let us assume that this question covers 10 chunks of knowledge. It could easily be, that in order to answer the question correctly, 5 chunks of information are needed. That is, those people who know 5 to 10 chunks will get the question correct while those who know less than 5 chunks will get the question wrong. Notice that the dichotomous question does not reflect the absolute variance in knowledge. Now, if at the beginning of the course the students have close to zero chunks of knowledge pertinent to this question, then the situation will tend to make the ratio of final knowledge much greater than the ratio of final test scores to initial test scores.

C. Differential Chunks Per Item. Consider now the situation where different items represent differing chunks of information. An item writer may write one question which covers 10 chunks of information while another item covers only one chunk. Yet correctly answering the latter is equal to answering the former on most every examination. Perhaps Table I with its corresponding Figure 3 will help present this situation. Notice that in this case I have let the number of chunks of information be perfectly correlated with the item difficulty. This would not always be the case, but it helps demonstrate the point. The subject with the least amount of knowledge would tend to get those questions which were of the lower difficulties. Thus, the treatment of items as equivalent on an exam when in fact they represent differing amounts of knowledge would tend to produce the curvilinear relationship as indicated in Figure 3, if those students with low amounts of knowledge tended to get correct those items which require small amounts of information. The test in this case represents a Guttman Scale. I would not contend that this type of scale would perfectly fit a test situation, but I would contend that it tends to lean in this direction.

TABLE I

CHUNKS OF KNOWLEDGE AND TEST SCORES
FOR A HYPOTHETICAL 100 ITEM TEST

No. of Questions	Item Difficulty	Chunks per Question	Test Score	No. of Chunks
10	1.0	1	10	10
10	.9	2	20	30
10	.8	3	30	60
10	.7	4	40	100
10	.6	5	50	150
10	.5	6	60	210
10	.4	7	70	280
10	.3	8	80	360
10	.2	9	90	450
10	.1	10	100	650

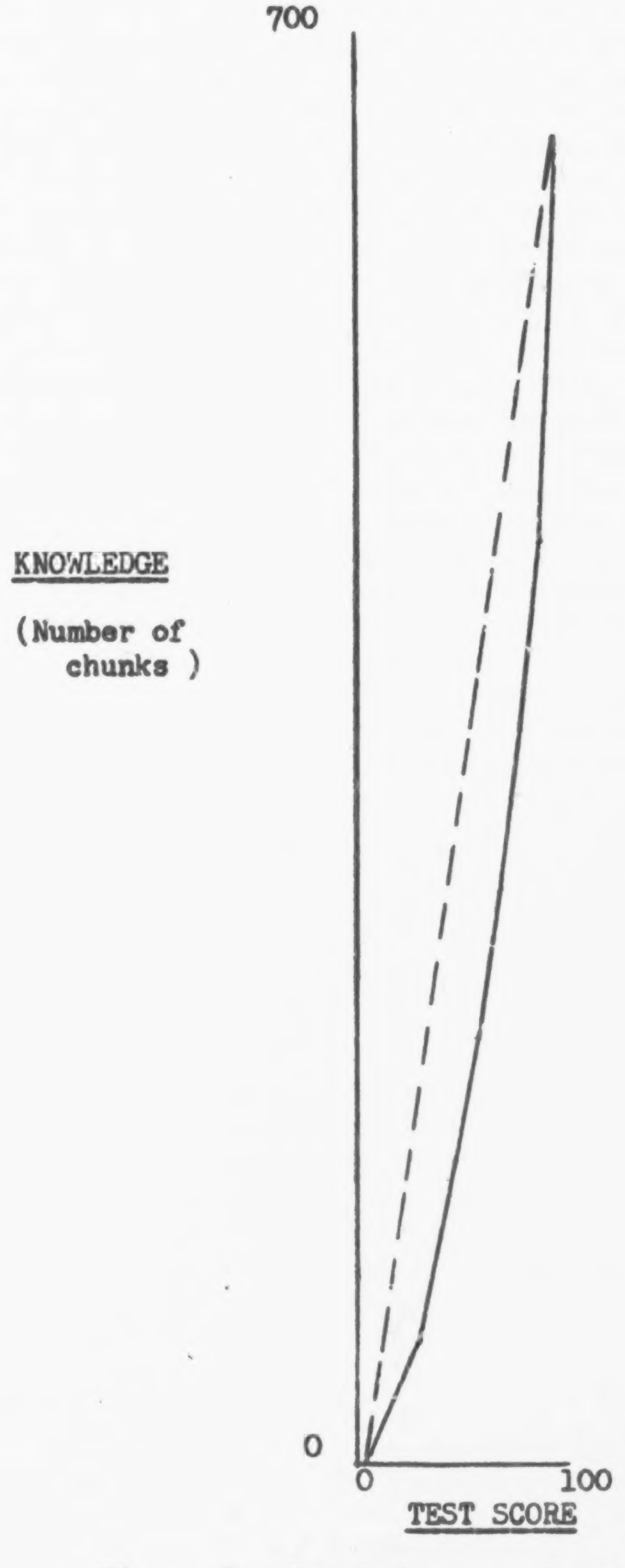


Figure 3. Knowledge as a Function of Test Score for Differential Chunks per Item

D. Chance Inequalities. Finally, we have the choice tests such as True-False, or Multiple Choice whereby one has only to pick an alternative answer, and thus, one can expect a certain number of correct answers by chance alone. Let us consider 5 questions, one each on the concepts: atom, molecule, neutron, proton, and electron. If each question concerned one of these concepts and the five alternatives for each question were all five of these concepts, then we would have a situation whereby a person with no knowledge would be expected to get one question correct by chance. However, a person who knew one concept would get that one correct, but he would not get 1/5 of those remaining questions. He would instead be expected to get in this case 1 + 1/4 of the remaining since he would not be guessing from 5 alternatives, but only 4 since he could definitely eliminate one of the alternatives. Thus, if one assumes that the per cent of the answers which are marked correctly by guessing is linearly correlated with the amount of knowledge, one will obtain the situation which is represented in Figure 4. In this case, the number of chunks of information is set at 100. From this 100 chunks, 20 were sampled and 5 choice (multiple-choice) items were written on these 20 chunks. The alternatives were assumed to be related to the other 80 chunks of information so that the more chunks one knew the more alternatives one would be able to eliminate. In this case the per cent of the questions which were answered correctly by chance was correlated perfectly with the amount of knowledge. This is the same as saying that the curve is constantly accelerated or is parabolic. The general equation for the curve is:

$$x = pn + \frac{2qn}{N} y - \frac{qn}{N^2} y^2$$

or

$$y = N - \frac{N}{\sqrt{qn}} \sqrt{n - x}$$

where x = the score on the test
y = the number of units of knowledge
n = the number of items on the test
N = the total number of units of knowledge
p = per cent of items expected to be correct by chance; unity divided by the number of alternatives
q = 1 - p

V. An Example of the Model in Action

Figure 5 presents an attempt to construct a hypothetical test situation. Curve 1 is the frequency distribution of the initial knowledge of 127 subjects. Curve 2 is the frequency distribution of final knowledge for the same 127 subjects. Curve 3 is the frequency distribution of initial test performance which was obtained from projections from Curve 1. This curve is the one that comes from the equations which I presented earlier. That is:

$$y = N - \frac{N}{\sqrt{qn}}\sqrt{n - x} \text{ or in this instance } y = 100 - 25\sqrt{20 - x}$$

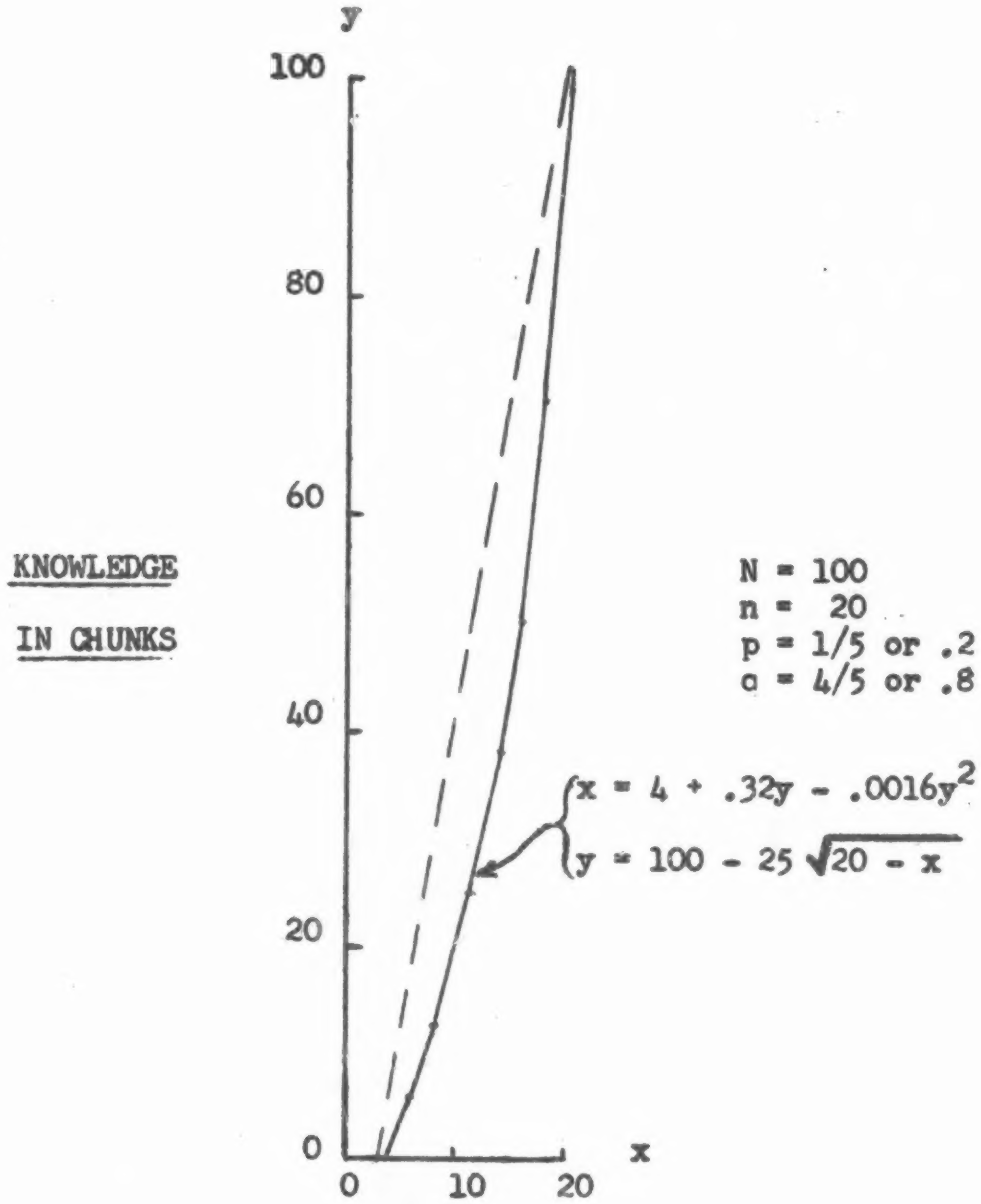


Figure 4. Knowledge as a function of Test Score when the chance p is correlated with Knowledge

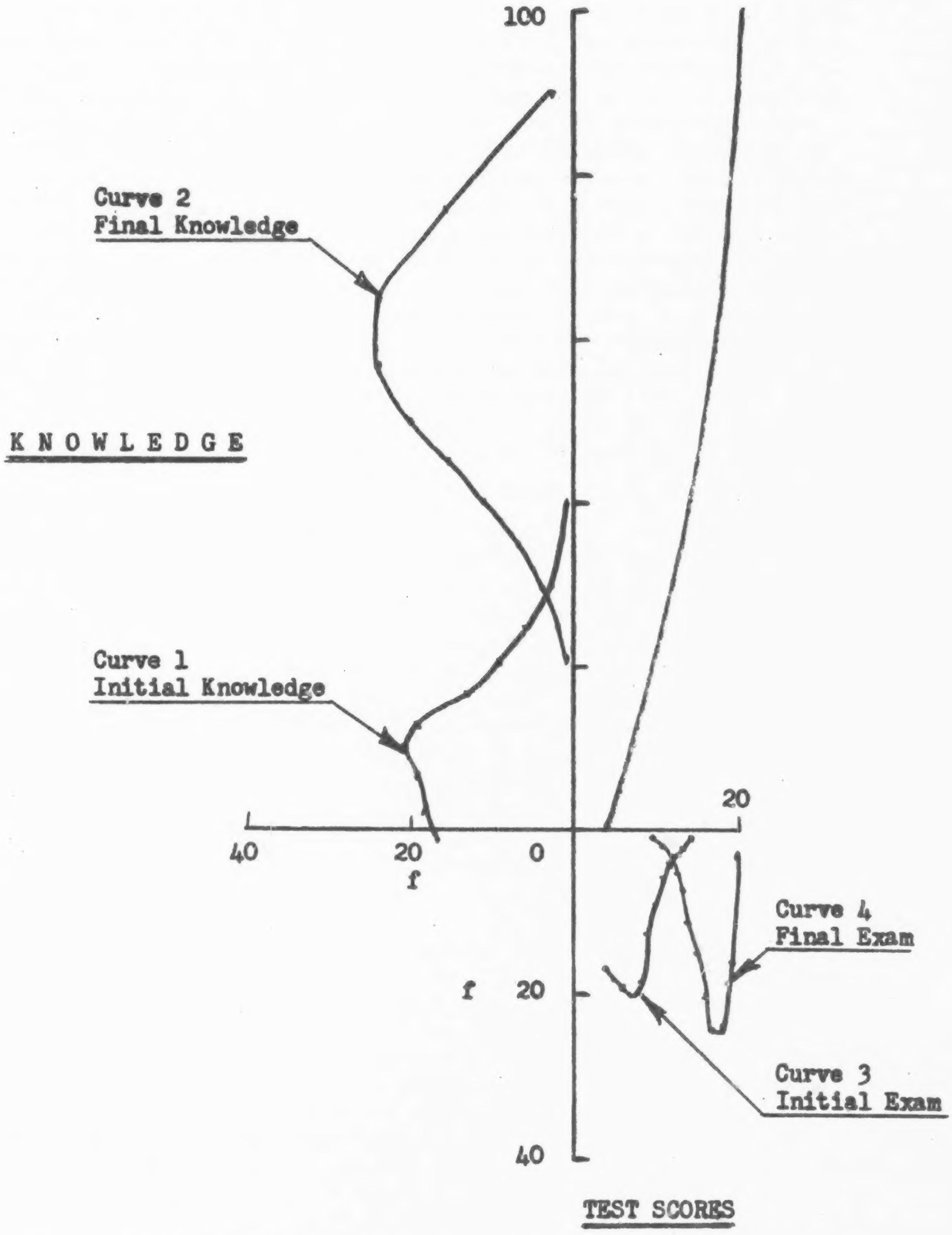


Figure 5. Frequency distribution of Knowledge and Test Scores on a hypothetical 20 Item Test. (N = 127)

Curve 4 was obtained in a similar fashion from Curve 2. The shape of the curves are like that of a binomial distribution with an N of 1,000. Curve 1 was generated with a p of .10 and Curve 2 was generated with a p of .60. Next, I constructed a situation whereby the correlation between initial knowledge (I) and final knowledge (F) was at its highest. This was accomplished by matching the lowest initial knowledge score with the lowest final knowledge score, and this matching continued on up the distribution until I had also matched the highest initial knowledge score with the highest final knowledge score. Since for each initial and final knowledge score there is a corresponding initial and final test score, one can see that by maximizing the correlation between initial and final knowledge one has at the same time maximized the correlation between initial and final test scores. Table II presents the intercorrelations between the four variables and the various gain scores. First note that the constructed correlation between initial and final knowledge (r_{IF}) was .94 while its counterpart for the test scores (r_{if}) was .92. The validity of the initial and final exams is very high (r_{iI} = 1.00 and r_{fF} = .98), yet the correlation between crude gain in test scores and crude gain in knowledge ($r_{(f-i)(F-I)}$) is only .24. The correlation between residual gain in test scores and crude gain in knowledge ($r_{(f.i)(F-I)}$) is much better, being .60. However, the highest correlate of crude gain in knowledge was that underrated indicant, "final exam." The correlation being a whopping big .92. Now before I continue, let me point out that the standard deviation of final knowledge was almost double that of initial knowledge while at the same time the standard deviation of final exam was slightly less than that of initial exam. Therefore it could well be that by simply looking at the variances of exams one has been lulled into thinking that they were good indicants of the variance of knowledge when it could be that they are not. The results can now be restated as follows: although initial and final exams may be very valid, and although the correlation between initial and final exam may be very high, and although one cannot assume zero or equal initial individual differences, it still may be that final exam is the best indicant of learning.

It is not always true that final exam is the best indicant of learning even within the confines of this model. From my brief exploration of the mathematical relationships I think one can generalize that as long as the ratio of final knowledge to initial knowledge is sufficiently greater than unity, then final exam will be the best indicant of learning, when the correlation between initial knowledge and learning remains moderately high. As the above ratio increases the size of the above correlation became less and less crucial. This means that it is my feeling that if a researcher can assume that the variance in final knowledge is much greater than the variance in initial knowledge, then he can safely use the final exam as the best indicant of learning. Also if one feels that he can safely assume a moderately high correlation between initial knowledge and learning then he would also seem to be secure in using the final exam as the best indicant of learning.

INTERCORRELATIONS AMONG THE VARIABLES CONSTRUCTED FROM THE HYPOTHETICAL TEST SITUATION IN FIG. 5

TABLE II

			(i)	(f)	(I)	(F)	(f-i)	(f.i)	(F-I)		
I			1	2	3	4	5	6	7	MEAN	S.D.
Initial Exam	(i)	1	1.00	.92	1.00	.95	-.41	.00	.76	7.2	2.32
Final Exam	(f)	2	.92	1.00	[illegible]	.98	-.02	.39	.93	16.3	2.13
Initial Knowledge	(I)	3	1.00	.90	1.00	.94	-.45	-.05	.73	11.0	8.56
Final Knowledge	(F)	4	.95	.98	.94	1.00	-.15	.26	.92	54.7	14.73
Crude Gain	(f-i)	5	-.41	-.02	-.45	-.15	1.00	.91	.24	9.1	0.91
Residual Gain	(f.i)	6	.00	.39	-.05	.26	.91	1.00	.60	---	---
Learning	(F-I)	7	.76	.93	.73	.92	.24	.60	1.00	43.7	7.27

VI. Suggestions for Research, Practical Usage, and Criterion Problems

It is my feeling that if an investigator attempts to measure classroom learning or to investigate the correlates of classroom learning he should take into account the fact that he is using only a rough instrument which only indicates relative amounts and is not an instrument which measures absolute amounts. Although most researchers realize this, I feel that they do not take into account the possibility that the assumption of a linear relationship between their indicant, and the variable in question may not be tenable.

Although I am sure that the curvilinear relationship which I have suggested will not be readily accepted, I do not feel that I should shoulder the entire weight of the responsibility for showing that the curvilinear relationship fits the situation best. After all, what evidence is there that the relationship is linear in the first place?

As for the criterion problem in learning, the topic for this conference, I feel that I should first state what I think the problem is. I think the problem lies in more adequately defining the variable which one is desirous of investigating. This includes an attempt to conceptualize the units of the variable learning. Then, the criterion problem in learning becomes either one of two things: (1) to measure the learning variable; or (2) to isolate the best indicant of that variable. My suggestions today have included a conceptual definition of classroom learning and a model which suggests the best indicant of that variable. On the other hand one could also mean something entirely different when one speaks of the criterion problem in learning. One could mean the problem of finding the best predictor of later success in some practical performance situation. In this case, the problem is not necessarily that of measuring or of defining learning, but becomes a problem of isolating that particular variable in the learning situation which correlates best with the performance variable. That variable may be final exam, final knowledge, learning, learning per time spent in study, or the ratio of final knowledge to initial knowledge. Thus the criterion problem defined in this manner would then involve a whole host of possible best indicants.

My suggestion to the conference would be that investigators should first attempt to define what they mean by classroom learning and then set about the task of establishing the relationship between learning and the indicants used, namely psychological tests. Once this relationship has been established the indicants can be calibrated and used to measure learning itself.

As the first step in the investigation of the problem, I would like to suggest that one might attempt to approximate final knowledge by using a curvilinear transformation on the final exam, or the initial exam if he has such information. One may find such a transformation would yield meaningful and useful results. Such a transformation is very simple with high speed electronic computors.

VII. Summary

Previous research using various measures of gain have often yielded results which are very contrary to logical expectations. It has long been recognized that this result could have come about because of our use of inadequate measures of learning. It was suggested in this paper that we are not so fortunate as to have inadequate measures of learning. We do not have any measures of learning at all. I have suggested that we may have, however, at our disposal some very good indicants or correlates of learning. I have suggested that test performance is curvilinearly related to knowledge and that if such is the case then final exam will most often be the best indicant of classroom learning.

It was also suggested that if a researcher desires to concern himself with classroom learning he need not limit himself to two alternatives, those being: (1) assume equal or zero initial individual differences or (2) administer some type of initial exam and to worry about how he should combine the two. I suggested that he has still a third alternative. He may assume that the variance of final knowledge is much greater than the variance of initial knowledge or he may assume that there is a significant positive correlation between initial and final knowledge. If he makes one or both of these assumptions then I have suggested that he may be justified in using the final exam by itself as the best indicant of learning.

Finally, I have suggested that if the criterion problem involves measuring learning then we should get started on the task of establishing the relationship between test scores and the learning variable so that we can then calibrate our test instruments. If however the criterion problem involves finding the best indicant of learning then we should concern ourselves with the task of finding that indicant which correlates highest with learning and not worry about calibration. It is my feeling that the most important task of research in classroom learning is to attempt to establish units of the learning variable itself so that the important task of calibrating our test instruments can be started. I do not think we have seriously considered the problem in this light, as yet. I think this is what S. S. Stevens would have us do if we want this area to become more mature as a science.

REFERENCES

Lord, F. M. Further Problems in the Measurement of Growth. Journal of Education and Psychological Measurement, 1958, 18, p 437-451.

Manning, W. H. and DuBois, P. H. Correlational Methods in Research on Human Learning. Perceptual and Motor Skills, 1962, 15, p 287-321.

Stevens, S. S. Measures and Indicants. Handbook of Experimental Psychology, S. S. Stevens (Ed.), John Wiley and Sons, New York, 1951, Chapter I, p 47-48.

THE CRITERION PROBLEM

Winton H. Manning

For so many years psychologists have been discussing the criterion problem that it seems difficult to say anything really new on the subject. To be sure, much past discussion is closely tied to research on test validation, where the obtaining of satisfactory measures of performance has been properly identified as the most difficult and fundamental problem in any selection research program. Robert Thorndike could not have put it much more strongly than when he said "This problem is absolutely central, for other research can hardly proceed until a criterion is provided, and the program of research can only be as good as that criterion." (Thorndike, 1949, p. 119). It seems inevitable, then, that we should from time to time return to discuss this question, even though, like the pilgrims who went to Canterbury, the principal benefits may perhaps be obtained in what transpires along the way, rather than in what may be found when we reach our destination.

There are typically two rhetorical stances one may take in approaching the problem of writing a paper on the criterion problem. The first is that of a "missionary" who devotes his efforts to showing that most of what we do, especially in measuring classroom learning, is wrongheaded, trivial, biased, and downright sinful in regard to selecting and constructing criteria of proficiency. After a long and detailed exposition of our shortcomings pointing especially to the discrepancies between curricula, goals, and evaluations of achievement a concluding exhortation is made to the effect that salvation may be attained only by finding "truly meaningful, relevant criteria." A second platform, often a reaction to this first, adopts a hard-headed position to the effect that we are overly introspective, indeed perhaps even morbid in ou analysis of the criterion problem. Rather, we should take a sophisticated approach in which we admit that the problem of finding truly meaningful, relevant criteria in the larger sense is not finally solvable. What we identify as the basis for our judgment is necessarily arbitrary, and is subject to criticism only on technical grounds, being justified frecuently on the basis that it constitutes a satisfactory operational definition of the concept in question.

The truth, I think, lies somewhere in between these two radical positions--neither the missionary nor the pragmatic sophisticate are correct in their appraisal of psychological research as regards the criterion problem. Nevertheless, it is incumbent upon me to make some suggestions of ways in which the criterion problem may be placed in a perspective that will permit greater generalizability and validity of our assessments of learning outcomes.

It will become evident as I develop the one or two modest points I wish to make that my thinking on the criterion problem has been considerably influenced by Campbell and Fiske (Campbell, 1954; Campbell and Fiske,

1959) who have emphasized the necessity of convergent and discriminant validation of tests by the means of a multitrait-multimethod matrix.

Historically, experimental psychology (especially in the field of learning), has emphasized the point that operational definitions occupy a central role in theory building. The essentially "literary" conceptions with which the psychologist may think, must be translated into an operational definition in terms of a test, a measuring instrument, or an objective behavioral record of some type. The behavioral record itself is further transformed into data as soon as we map it onto some kind of measuring scale. Thus, as Coombs (1963) has pointed out the experimenter actively engages in a process in which he first selects only some of the behavior as being relevant to his concept and secondly, transforms this record into a qualitative or quantitative expression which is still further abstracted from the original behavior. Such a process of selection and transformation is risky in the sense that it may lose the important and retain the trivial. On the other hand, as we have seen again and again, these processes of abstraction and transformation may be the only means by which we can see clearly what is important and relevant in the situation. Essentially, this is the model which has made physics a success and we have no doubt emulated it for that reason. The movement from concept to operation, then, is a move which all hard-headed behavioristically oriented psychologists applaud, and it is the direction of movement which is most discussed in the literature of psychology. What about, however, the movement from operation to construct? Probably this is one of the sources of the great concern manifested about criteria in learning, especially in classroom research. What, for example, is the G.P.A. measuring? How can we make the move from operation to construct validly? First, of course, we must dispense with an overly restrictive definition of operationism. Years ago, Bridgman (1927, p. 10) pointed out that "if we have more than one set of operations, we have more than one concept, and strictly, there should be a separate name to correspond to each set of operations." Such a view must be rejected, in my judgment, as far too narrow and constraining. Rather, we must find a ground upon which we may discover converging clusters of measures, rather than seeking complete congruence. Such a convergent operationism emphasizes as equally important the "operation to construct transformation" and its converse. (Campbell and Fiske, 1959)

In this light, let us consider briefly a study conducted by Roff, Payne, and Moore (1954) which dealt with the analysis of a large number of parameters of motor learning. Of the 52 variables which were measured, 39 were derived from the learning curves of 175 airmen in three psychomotor tasks: -- the Complex Coordinator, the Rotary Pursuit, and the Multidimensional Pursuit. The 13 learning variables were essentially the same for all three tasks:

(a) sum of all trials or total score
(b) sum of first three trials
(c) sum of middle three trials
(d) sum of last three trials
(e) average slope
(f) ratio of early to late slope
(g) slope at y = 1 (initial learning rate)
(h) slope at y = 15 (intermediate learning rate)
(i) slope at y = 40 (terminal learning rate)
(j) difference between first three trials and last three trials
(k) square of difference between successive trials (performance variability)
(l) variance of raw scores for the 40 trials around the mean
(m) fluctuation function of raw scores around individual cumulative mean curve.

In addition, a battery of 13 paper and pencil aptitude and ability tests was also administered. The entire set of variables was then subjected to an oblique multiple group factor analysis with the result that 16 factors were defined. These were a performance factor for each of the psychomotor tasks, an early slope factor for each of the tasks, a variability factor for each of the tasks, a late slope factor for each of the tasks, and four ability factors: verbal, numerical, mechanical, perceptual-spatial. The important thing to note is that each of the factors is method and task specific, and further that each is independent of the ability measures. This is not unlike the results of Anderson, as reported by Campbell and Fiske (1959) in which measures of hunger, thirst, and sex drives were more highly correlated within the obstruction box or activity wheel method than were the same drives across methods or apparatus. Similar results have been widely reported and widely repressed, for the high proportion of methods variance makes quite obvious one of the sources of difficulty in moving from operation to construct in dealing with learning criteria.

Herein, it seems, must lie both the source of the criterion problem as well as the framework for handling it. There is no way that I know of by which we may extract from the variability of a variable the proportion of variance associated with the method and with the task, unless we have convergent information about the performances of Ss in a variety of tasks using a variety of methods. Such a multitask, multimethod, multivariate approach to the problem of criterion development will require a much larger investment of effort than is normally undertaken, but less than this, it seems, may be fruitless.

A serious consideration of this proposal leads to an attempt to formulate a kind of framework by which we may seek to understand the nature of the criterion measures we employ, and their generalizability across samples and across situations or contexts. Furthermore, it would have the effect, possibly, of pointing up the extent to which research designs have become routinized and stereotyped, to the detriment of our understanding and scientific productiveness. Let me invite your attention to Figure 1. Three axes are represented, each indicating an important way of viewing experimental approaches to the study of learning, especially perhaps human learning. These dimensions are:

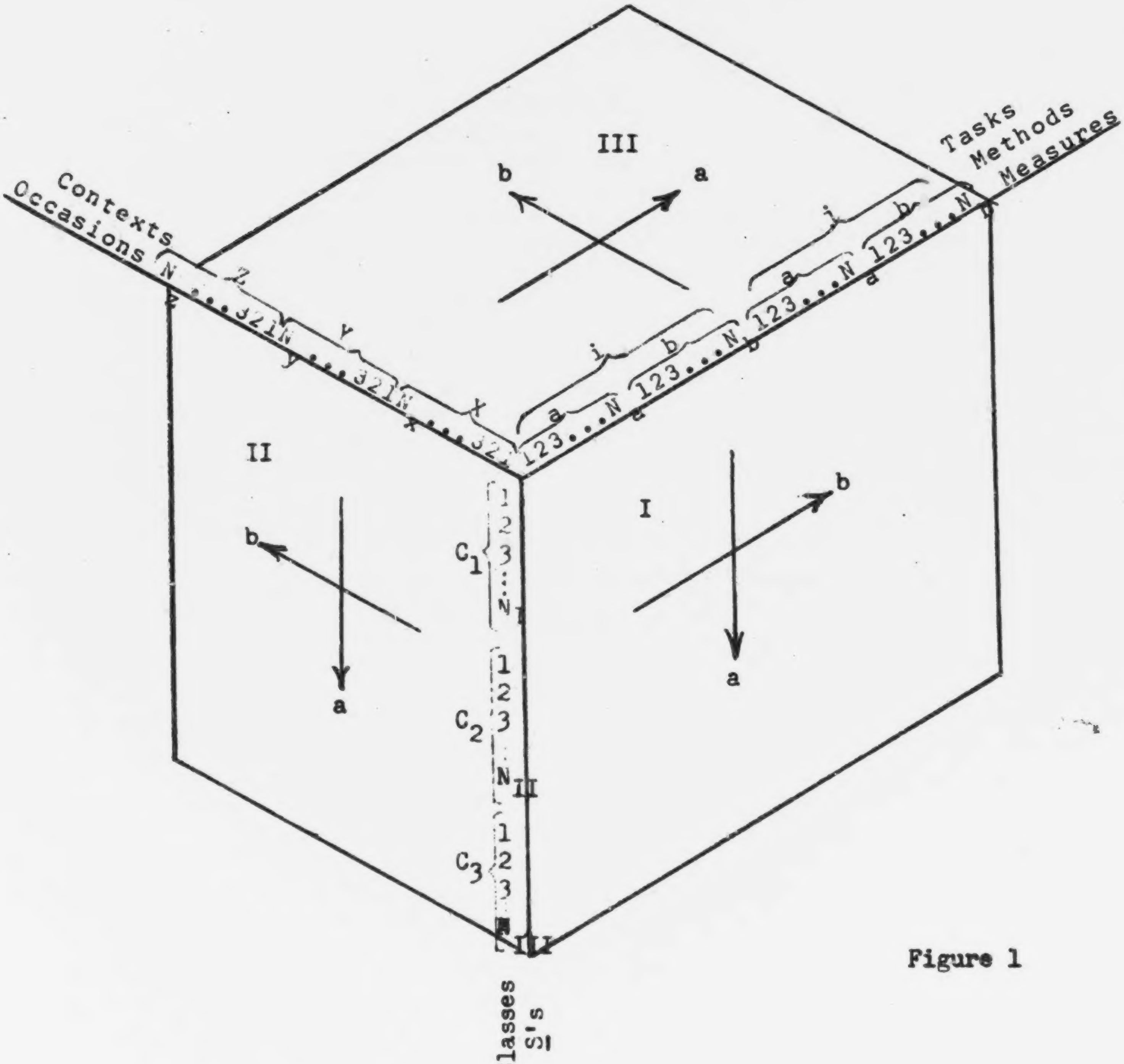

Contexts
Occasions
Tasks
Methods
Measures
III
II
I
a
b
N ...321 N ...321 N ...321
z
y
x
123...N 123...N 123...N 123...N
a
b
i
j
C_1
C_2
C_3
1 2 3 ... N
N_I
N_{II}
N_{III}
Classes
$\underline{S}$'s

Figure 1

Data Class	Relationships between	Summed over	Constant Condition
Ia	Tasks-methods-measures	persons or classes	single or pooled occasions or contexts
IIIb	Tasks-methods-measures	occasions or contexts	single subject or class of subjects
IIa	Contexts or occasions	persons or classes	single task, method, or measure
IIIa	Contexts or occasions	tasks, methods, measures	single or pooled persons or classes
Ib	Persons or classes	tasks, methods, measures	single or pooled contexts or occasions
IIb	Persons or classes	contexts or occasions	tasks, method, or measure

1. A task-method-measure dimension.
2. A subject-class dimension.
3. An occasions-context dimension.

It is obvious that I have combined certain ideas in a nested fashion so as to make more evident salient aspects of the problem. Another schema for another purpose might produce a different configuration or greater utility (Gulliksen, 1958; Cattell, 1952).

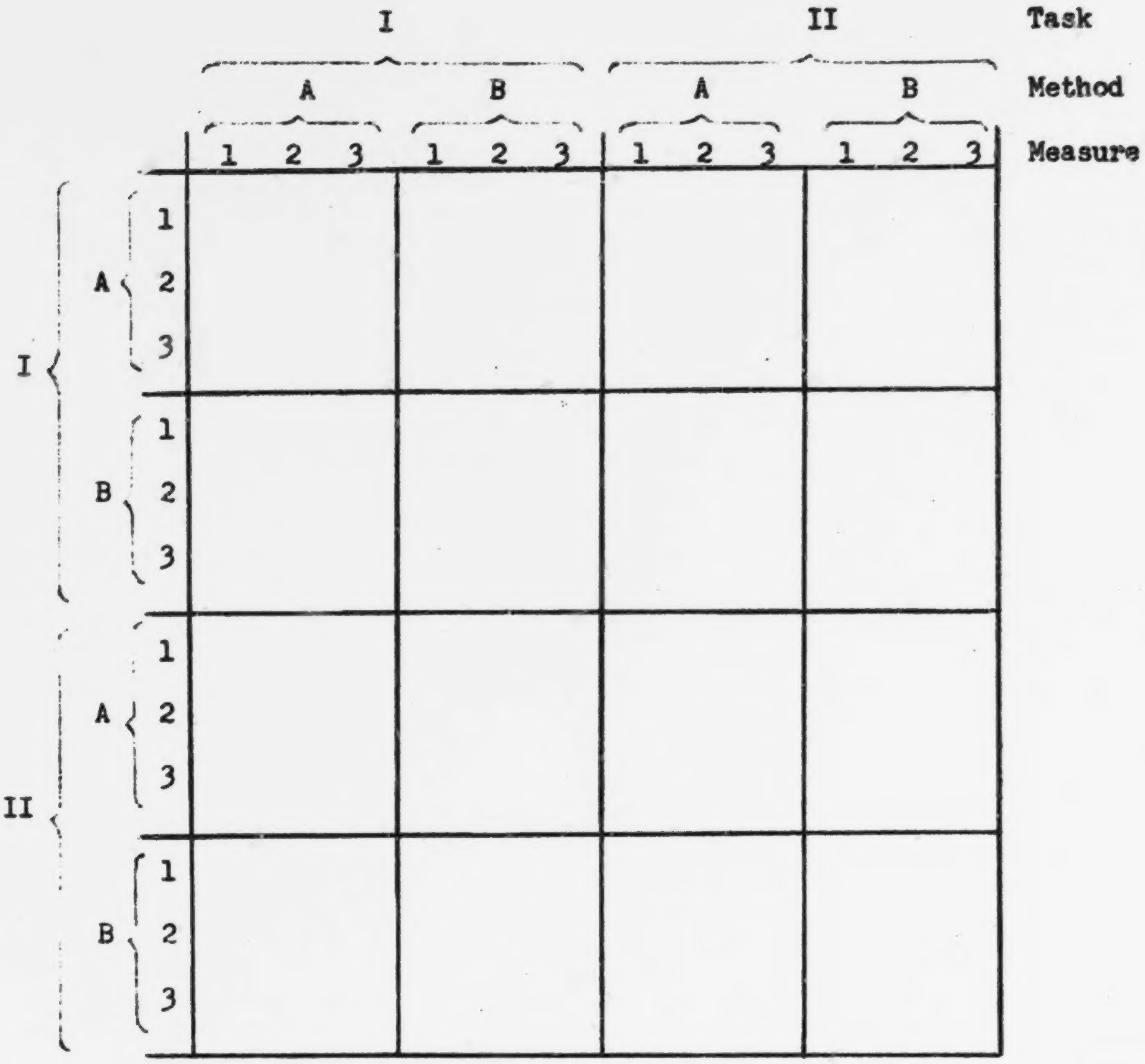
One of the implications of this configuration is that we may describe at least six classes of research problems corresponding to the two surfaces which intersect in each dimension. Let me invite you to turn now to Figure 2 which deals with what seems historically the most important problem for criterion development, namely the "task-method-measure dimension."

I. Task-method-measure dimension:

A consideration of the task-method-measure matrix in Figure 2 suggests two kinds of data of interest. Firstly, we may speak of the data generated by an obtaining relationship between task-method-measures for a particular occasion or context, and summing over a number of persons. This is referred to as data Ia. On the other hand, we could obtain the interrelationships of a task-method-measure variable, for a particular person or class of persons over a number of occasions and contexts. This would be data IIIb. The second type of data is less frequently obtained, but is relevant, for example, to the question of whether the course of improvement or susceptibility to change as a function of contextual differences is parallel or similar for different tasks,methods, or measures for a particular class of learners. This is a second way of investigating the similarity structure, so to speak, of different criteria of learning, for by this means we may see how these variables behave similarly over time. Furthermore, we may compare individuals or groups of subjects, since such an approach implies that points along this latter dimension are parameters.

Although the classroom situation may present itself as an obvious example of data Ia let us consider a case in rote learning as provided in the work of Robert Stake (1958). In Tables 1 and 1a are found the intercorrelations of three parameters of the learning curves of 240 children for three rote memory tasks. Task I involves the matching of a stimulus word displayed in a window to a response word found on a switch panel. Task II is also individually administered, and involves writing down of as many of a list of 16 verbs as he can in any order, after hearing these read. Task III is the same except that the task is administered in a group and the words are adjectives rather than verbs.

In this example task and method are not separable, but the picture presented is probably illustrative of the general situation in regard to learning measures. Measures are to a considerable degree both task specific and measurement specific, in ways which are not wholly predictable from the logical relations of the tasks or measures. It is evident from these data, however, that Tasks II and III are most closely related and that although the error and curvature parameters have a moderate


I
II
Task
A
B
A
B
Method
1 2 3
1 2 3
1 2 3
1 2 3
Measure
I
A
1
2
3
B
1
2
3
II
A
1
2
3
B
1
2
3

Ia. Relations between task-method-measures summed over persons for a given occasion or context.

IIIb. Relations between task-method-measures summed over occasions or contexts for an averaged (rarely single) person or class of persons.

Figure 2.

Table 1. Intercorrelations of three parameters for three tasks (from Stake, 1958).

		I (Word Match)			II (Word Memory 1)			III (Word Memory 2)		
		E (error)	C (curve)	F (fit)	E (error)	C (curve)	F (fit)	E (error)	C (curve)	F (fit)
	E	-	47	15	36	25	00	42	17	15
I	C	47	-	19	23	15	-03	29	24	-06
	F	15	19	-	02	02	02	07	06	-06
	E	36	23	02	-	89	-22	66	42	14
II	C	25	15	02	89	-	-20	48	40	-06
	F	00	-03	02	-22	-20	-	14	16	15
	E	42	29	07	66	48	14	-	74	-11
III	C	17	24	06	42	40	16	74	-	-15
	F	15	-06	-06	14	-06	15	-11	-15	-

Table 1a. Rearranged data from Table 1 above.

		Error			Curve			Fit		
		I	II	III	I	II	III	I	II	III
	I	-	36	42	47	25	17	15	00	15
E	II	36	-	66	23	89	42	02	-22	14
	III	42	66	-	29	48	74	07	14	-11
	I	47	23	29	-	15	24	19	-03	-06
C	II	25	89	48	15	-	40	02	-20	-06
	III	17	42	74	24	40	-	06	16	-15
	I	15	02	07	19	02	06	-	02	-06
F	II	00	-22	14	-03	-20	16	02	-	15
	III	15	14	-11	-06	-06	-15	-06	15	-

relationship the "fit" parameter is almost wholly independent of either of these. Numerous other examples of this same kind of situation could be supplied. What these data tell us is that we must treat criteria in a learning situation by means of a convergent operationism if we are to develop constructs of wide generalizability. Otherwise, we are likely to make quite erroneous judgments.

I shall not dwell on the remaining data classes at any great length except to point out some of the possibilities inherent in them in regard to the criterion problem.

II. Contexts-occasions dimension:

Data IIa and IIIa. A study of the relationship of occasions to one another, for a particular task summed over subjects is well illustrated by the work of Fleishman (1955) as well as the many other studies seeking to identify the factors associated with changes in performance in a particular task. The recent work of Ledyard Tucker (1959) reported at an earlier conference in this series is also relevant. In Tucker's work two simple probabilistic learning tasks were examined and it was shown that in the one, a single latent underlying learning curve was sufficient, but, in the other, three such latent curves were implied by the data. Presumably criteria of proficiency should reflect these underlying properties or factors, and by systematically applying Tucker's analysis to a logically related set of tasks it is possible that we might learn a great deal about the nature of task complexity and how it changes with practice.

III. Person-classes dimension:

Data Ib and IIb. In considering the nature of the learner the work of Tucker mentioned earlier is also particularly relevant. As Tucker pointed out some of the persons in his study seemed in the more complex task to be learning Factor A, some others Factor B, and still others Factor C. If we could develop discriminant functions by which to assign subjects to common classes or groups based upon a number of tasks rather than only one, it would perhaps be possible to match subjects to learning conditions and methods more intelligently.

Even when we deal with a very simple task, classes of subjects may not learn at the same rate. For example, in one study done as a master's thesis by McLean (1959) in our laboratory ten groups of Ss stratified on the basis of initial performance and they were compared using the first and second derivatives of the smoothed learning curves as criteria. The task was learning to print the alphabet upside down. Highly significant differences existed among these groups in both respects. Similar data have been reported by Reynolds and Adams (1953) for the Rotary Pursuit Test, although some contradictory results are also found in the literature. Whether it is possible to partition Ss into classes based solely upon aptitude and ability tests is in my judgment not likely, for it appears that a significant portion of the variability of subjects in learning situations is not measurable by such tests.

Selection of particular ways of optimally assigning subjects to tasks and to contexts will be feasible in those situations where a great deal of convergent information about the relations of these tasks and contexts to one another has been generated. It simply appears unlikely to me that a particular set of tests will have very wide generality for predicting learning criteria but rather that only as we try to build into the test some of the same sources of task and contextual variance as are found in the criteria will we have markedly improved success in this respect.

In the foregoing discussion we have indicated some of the types of data classes and the ways in which the criterion problem relates to them. A further implication of this line of thought is that of developing, with reference at least to the first dimension of tasks-methods-measures, a taxonomic approach to the description of the learning situation. Such a description might take the form of a hierarchical factor or cluster analysis in which variance associated with a particular measure would be partitioned into common factors, factors associated with methods, factors associated with task contents, factors associated with measures, and, of course, unique variance. For example, in a classroom situation we might speak of:

(a) common factors underlying general achievement in a broad spectrum of learning activities;

(b) factors associated with particular tasks; say different subject matter areas, such as mathematics, English, reading, and foreign languages;

(c) factors associated with methods; say objective tests, peer ratings, self ratings, teacher ratings, essay tests, and so forth; and

(d) factors associated with measures--such as final proficiency, improvement in proficiency, and retention or relearning measures.

Such an approach would also suggest that further research using factorial analysis of variance designs permitting testing of differences and interactions in a tasks X methods of assessment X measurement variables X subjects design. Much more work of a descriptive nature probably needs to be done first, however, before analyses of variance are employed.

It would be a mistake to suppose that very much could be accomplished by submitting a haphazard collection of tasks, measures and methods to such an analysis. However, by carefully defining first the domain in which we are interested, fruitful results might be obtained. Consider, for example, one of the schools of the Naval Air Technical Training Command. Could we not find portions of the existing curriculum of a school in which we could identify say four kinds of tasks: (1) a verbal rate task, (2) a non-verbal rate task, (3) a verbal relational task, (4) a non-verbal or performance relational task. Methods of assessment might include objective test scores, ratings by fellow students, ratings by superiors, and evaluations of attainment by means of programmed instructional devices. Measures could include achievement in school, rate or change measures in school, and retention measures. This would alone produce a total of 48 variables, the analysis of which by the means we have briefly implied would be likely to be productive in assessing the criteria of learning in these schools. Extension of this approach to operational fleet contexts would also be possible.

Summary:

In summary, we are suggesting that the criterion problem stems at least in part from our difficulty in generalizing from the available operational measures to meaningful constructs. Our principal difficulty, it seems to me, is not that we are doing the wrong things, but that we are not doing enough. An extension of the Campbell and Fiske approach to thoroughgoing analysis of our criterion data within a multitask-multimethod-multimeasure matrix offers promise as a means of developing convergent understanding of the criterion structure. Nothing inherent in the nature of the problem would prevent us from extending this framework to include a temporal dimension reflecting occasions and contexts or an individual differences dimension comprising subjects, classes, or species of subjects. In placing our emphasis on such a multidimensional matrix format, we have not meant to suggest that simply multiplying our measures will assist us. On the contrary, as measures, task, and method are increased an increased burden of logical, relational constraints should be placed upon us. Otherwise it appears unlikely that we shall do much more than throw sand in our eyes. Nevertheless, it is sobering to consider that perhaps even the simplest learning situation is too complex for such an undertaking. In such a debate, however, I would vote against the Walrus and the Carpenter, when as the Reverend Charles Lutwidge Dodgson (1870) said (while wearing his other hat):

The Walrus and the Carpenter
Were walking close at hand;
They wept like anything to see
Such quantities of sand:
"If this were only cleared away"
They said, "it would be grand!"

"If seven maids with seven mops
Swept it for half a year,
Do you suppose," the Walrus said,
"That they could get it clear?"
"I doubt it," said the Carpenter,
and shed a bitter tear.

References

Bridgman, P. W. *The Logic of Modern Physics*. New York: MacMillan, 1927.

Campbell, D. T. "Operational delineation of 'what is learned' via the transposition experiment." *Psychological Review*, 1954, 61, 167-174.

Campbell, D. T., & Fiske, D. W. "Convergent and discriminant validation by the multitrait-multimethod matrix." *Psychological Bulletin*, 1959, 56, 81-105.

Cattell, R. B. *Factor Analysis*. New York: Harper, 1952, 1-462.

Coombs, C. *A Theory of Data*. New York: Wiley, 1963.

Fleishman, E. A., & Hempel, W. E. "The relation between abilities and improvement with practice in a visual discrimination reaction task." *Journal of Experimental Psychology*, 1955, 49, 301-312.

Gulliksen, H. *Mathematical Solutions for Psychological Problems*. Technical Report on ONR Research Contract Nonr 1858(15) and NSF Grant G-642, Princeton University and Educational Testing Service, 1958, 1-54.

Hodgson, C. L. (Lewis Carroll) "The Walrus and the Carpenter" from *Through the Looking Glass*, 1870.

Manning, W. H., & DuBois, P. H. "Gain in proficiency as a criterion in test validation." *Journal of Applied Psychology*, 1958, 42, 192-194.

Manning, W. H.,& DuBois, P. H. "Correlational methods in research on human learning." *Percept. mot. Skills*, Mon. Suppl. 3VI5, 1962, 15, 287-321.

McLean, M. L. *Characteristics of the Learning Curve as a Function of Initial Performance*. Unpublished M.A. Theses, Texas Christian University, 1959.

Reynolds, B., & Adama, J. A. "Psychomotor performance as a function of initial level of ability." HRRC Resch. Bull., 53-59, October 1953.

Roff, M., Payne, R. B., & Moore, E. W. "*A Statistical Analysis of the Parameters of Motor Learning*. USAF School of Aviation Medicine, Project No. 21-0202-0001, Report No. 1, February 1954, 1-22.

Stake, R. *Learning Parameters, Aptitudes, and Achievements*. Technical Report, ONR Research Contract Nonr 1858(15) and NSF Grant G-642. Princeton University and Educational Testing Service, 1958.

Thorndike, R. L. *Personnel Selection*. New York: Wiley, 1949.

Tucker, L. "Determination of Generalized Learning Curves by Factor Analysis," in *Factor Analysis and Related Techniques in the Study of Learning*. Edited by P. H. DuBois, W. H. Manning, & C. J. Spies, Technical Report No. 7, ONR Contract Nonr 816(02), August 1959, 143-168.

PROGRAMMED INSTRUCTION AND THE ABILITY TO LEARN

James L. Wardrop

This study was carried out in order to investigate the use of programmed instruction as a miniature learning situation for predicting performance in a subsequent large-scale (classroom) learning situation.

There are five assumptions or findings on which this study is based.

First, intelligence and the ability to learn are apparently not entirely the same. As early as 1901, Wissler (1901) found little or no relationship between scholastic achievement and a number of "mental tests" developed by Cattell. In more recent years, studies by Woodrow (1938, 1946) and Simrall (1947) indicate that intelligence and learning are not the same, and that the factors or abilities measured by intelligence tests are only partially those involved in the learning process. In view of the findings such as these, it should be possible to improve upon the use of intelligence tests for selection and prediction-particularly in educational situations (cf. Sorenson, 1963).

Secondly, the ability to learn seems not to be a unitary function. The findings of Wimms (1907), in one of the earliest studies in this area, set the pattern. He found no correlation between gain in two similar learning tasks. The majority of subsequent studies in this area report a similar lack of general learning ability (Allison, 1960; Atkinson, 1929; Hall, 1936; Husband, 1939; Stake, 1961).

In the third place, a potentially valuable approach to the prediction of learning is through the use of learning tests, or miniature learning situations. Such an approach, suggested by Fredriksen et al. (1947), has been used in a number of investigations of learning and human ability (see Allison, 1954; Allison, 1956; Allison, 1960).

Fourth, a major problem in the investigation of learning has been the statistic of measurement of learning. Most studies in the literature make use of one of three different measures: measures based on a single evaluation of proficiency at the conclusion of practice or training; crude gain, the difference between measures of final and initial proficiency; and percent gain, usually defined as the ratio of crude gain to initial status. Two other measures of "learning" have appeared in the literature in recent years: parameters of individual learning curves, involving curve-fitting and the determination of certain parameters of the curves so obtained; and residual gain, defined as that portion of the measure of final performance which is statistically independent of initial status. Because it offers the advantages of consistency, adaptability, and statistical logic, residual gain is the measure employed in this study.

Finally, programmed instruction provides a controlled, organized miniature learning situation amenable to careful, periodic assessment of the progress of learning, particularly initial and final proficiency. According to Green (1962, p. 112), the "learning process as it is controlled by programmed instruction differs in no essential way from the learning process as it is controlled in the classroom." The particular idea of using teaching machine performance to develop a measure of learning ability has been proposed by Sorenson (1963).

METHOD

A. Subjects

The subjects used in this study were trainees at the Naval Air Station, Memphis, Tennessee. Two groups were used: 148 students in the Aviation Mechanical Fundamentals School [AMFU(A)], and 330 in the Aviation Electronics Fundamentals School [AFU(A)]. The two groups were tested together in groups of about 45 over a period of eleven weeks, before they began school. Average age of the subjects was 19 years, and the average educational level was 12 years. The AFU(A) students averaged about one standard deviation above the means on the subtests of the Navy Basic Test Battery, a test of general intelligence, while students in AMFU(A) school averaged only one or two points above the means on this battery.

B. Tests and Procedure

The General Classification subtest of the Basic Test Battery, a group test of verbal intelligence, was used as the measure of intelligence. The learning tests used were the DuBois-Bunch Learning Test, a simple perceptual learning task adapted for group administration (DuBois & Bunch, 1949); and a Numbers Test, in which subjects were to trace, in sequence, numbers from 1 through 60 printed in apparently random positions on a page. (A fuller description of this test can be found in Hackett, 1964.) Each of these tests consisted of ten trials, from which an initial score (average score on the first two trials) and a final score (average score on the two final trials) were obtained.

The programmed instruction learning measure was obtained from performance on an 85 minute linear program on study skills. Before this 214-frame program, subjects were given a 27-item pretest over the material in the program; after completing the program, they were given an equivalent 27-item post-test over the same material. Included in the program were a number of "test" items, frames in which no answers were supplied and the students were required to write their responses.

The other test given during this pre-school session was a pretest over the material taught in the two fundamentals schools. The items on this pretest were taken from a pool of items used in the preparation of examinations in the schools. An effort was made to select items which applied to both schools. This test was used as a measure of initial knowledge or proficiency.

Measures of final status were not the same for the two schools. For AFU(A) students, the final average in the first (five-week) phase of the 19-week school was used, while the school final average was used for the AMFU(A) students, since this is only a five-week course.

Testing sessions were held once weekly, and each lasted approximately three hours. The learning tests, the How-to-Study program, and the school pretest were given in this session.

In the data analysis, a matrix of intercorrelations of all variables in the study was obtained for each school. From these matrices, the correlations of the residual gain measures with all other variables were found. Finally, intercorrelations among these residuals were obtained. The residuals used were gain ("learning") on the DuBois-Bunch Learning Test, learning on the Numbers Test, learning by means of programmed instruction, and learning in the classroom.

RESULTS

Table I shows the correlations of the predictors with two criteria, phase (or school) final average and a gain measure of classroom learning. In AFU(A) school, the programmed instruction learning measure correlated .30 with phase one final average, and .27 with the gain measure of classroom learning. These were significantly greater than the correlations of the other learning measures with either criterion. The intelligence measure (GCT) correlated .35 with final average and .28 with the gain measure of classroom learning. In AMFU(A) school, the programmed instruction learning measure correlated .25 with school final average and .23 with the gain measure of classroom learning. Again, these correlations were significantly greater than those involving the other learning tests. The intelligence measure (GCT) correlated .33 with final average and .28 with the gain measure in this school.

DISCUSSION

The results indicate that programmed instruction is more closely related to classroom learning than are the other learning tests employed. (It should be noted that neither of the other tests is a verbal learning test; rather, both are tests involving perceptual-motor skills.) In addition, when considered in combination with the intelligence measure, the programmed instruction learning measure raises the correlation with classroom performance (final average) from .33 to .34 in AMFU(A) school and from .35 to .38 in AFU(A) school. On the basis of these findings, several further studies are planned to determine more precisely the value of a programmed learning task in the prediction of classroom performance.

ACKNOWLEDGMENTS

The study reported here was carried out under Contract Nonr 816(14) between the Office of Naval Research and Washington University. The Computer Center at Washington University was responsible for statistical analysis of the raw data and generation of the correlation matrices.

Table I

Predictor-Criterion Correlations

	AFU(A) School	
	Final Average	Classroom Gain
GCT	.35	.28
Gain on DuBois-Bunch Test	.12	.08
Gain on Numbers Test	.16	-.02
Gain on Programmed Instruction	.30	.27

	AMFU(A) School	
	Final Average	Classroom Gain
GCT	.33	.30
Gain on DuBois-Bunch Test	.11	.13
Gain on Numbers Test	.04	.01
Gain on Programmed Instruction	.25	.23

References

Allison, R. B. Learning measures as predictors of success in Torpedoman's Mates School. Office of Naval Research Technical Report. Princeton, N. J.: Educational Testing Service, 1954.

Allison, R. B. Learning measures as predictors of success in Pipefitter and Metalsmith Schools. Office of Naval Research Technical Report. Princeton, N. J.: Educational Testing Service, 1956.

Allison, R. B. Learning parameters and human abilities. Office of Naval Research Technical Report. Princeton, N. J.: Educational Testing Service, 1960.

Atkinson, W. R. The relation of intelligence and of mechanical speeds to the various stages of learning. Journal of experimental Psychology, 1929, 12, 89-112.

Frederiksen, N., Carstater, E. D., & Stuit, D. B. Problems for further study. In D. B. Stuit (Ed.), Personnel research and test development in the Bureau of Naval Personnel. Princeton, N. J.: Princeton University Press, 1947.

Hackett, E. V. The use of learning tests in the analysis of training. In P. H. DuBois and K. M. Wientge (Eds.), Strategies of research on learning in educational settings. St. Louis: Washington University, 1964.

Hall, C. S. Intercorrelations of measures of human learning. Psychological Review, 1936, 43, 179-195.

Husband, R. W. Intercorrelations among learning abilities: I. Journal of Genetic Psychology, 1939, 55, 353-364.

Simrall, D. Intelligence and the ability to learn. Journal of Psychology, 1947, 23, 27-43.

Sorenson, A. G. The use of teaching machines in developing an alternative to the concept of intelligence. Educational and Psychological Measurement, 1963, 23, 323-329.

Stake, R. E. Learning parameters, aptitudes, and achievements. Psychometric Monographs, 1961, No. 9.

Wimms, J. H. The relative effects of fatigue and practice produced by different kinds of mental work. British Journal of Psychology, 1907, 2, 153-195.

Wissler, C. The correlation of mental and physical tests. Psychological Review Monograph Supplement, 1901, 3, No. 6.

Woodrow, H. The relation between abilities and improvement with practice. Journal of Educational Psychology, 1938, 29, 215-230.

Woodrow, H. The ability to learn. Psychological Review, 1946, 53, 147-158.

UNIVERSITY COLLEGE RESEARCH PUBLICATIONS

1. Strengthening the Residential Adult Education Experience. Sponsored by the Conference and Institute Division of the National University Extension Association, The Center for the Study of Liberal Education for Adults, and University College, November 1962, 66 pp.

2. Objectives and Methods of Research in Adult Education. Edited by Philip H. DuBois and King M. Wientge, May 1962, 51 pp.

3. Studies of Biographical Data in Adult Education. Philip H. DuBois and King M. Wientge, August 1963, 8 pp.

4. Strategies of Research on Learning in Educational Settings. Edited by Philip H. DuBois and King M. Wientge, February 1964, 74 pp.

5. Factors Associated with the Achievement of Adult Students. King M. Wientge and Philip H. DuBois, 1964, 39 pp.

6. Survey of Tuition Aid Plans of Business, Industry, and Government in Metropolitan St. Louis. King M. Wientge and Malcolm Van Deursen, June 1965, 12 pp.

7. Workshop for Counselors and Educators Concerned with the Education, Training and Employment of Minority Youth, Final Report, Part I: Development, Program, Evaluation. King M. Wientge, October 1965, 71 pp.

8. Workshop for Counselors and Educators Concerned with the Education, Training and Employment of Minority Youth, Final Report, Part II: Discussion Guide to the Problems of the Culturally Deprived: An Introduction for Teachers and Counselors. John M. Whiteley and King M. Wientge, November 1965, 71 pp.

9. Criteria in Learning Research. Edited by King M. Wientge and Philip H. DuBois, 1966, 65 pp.